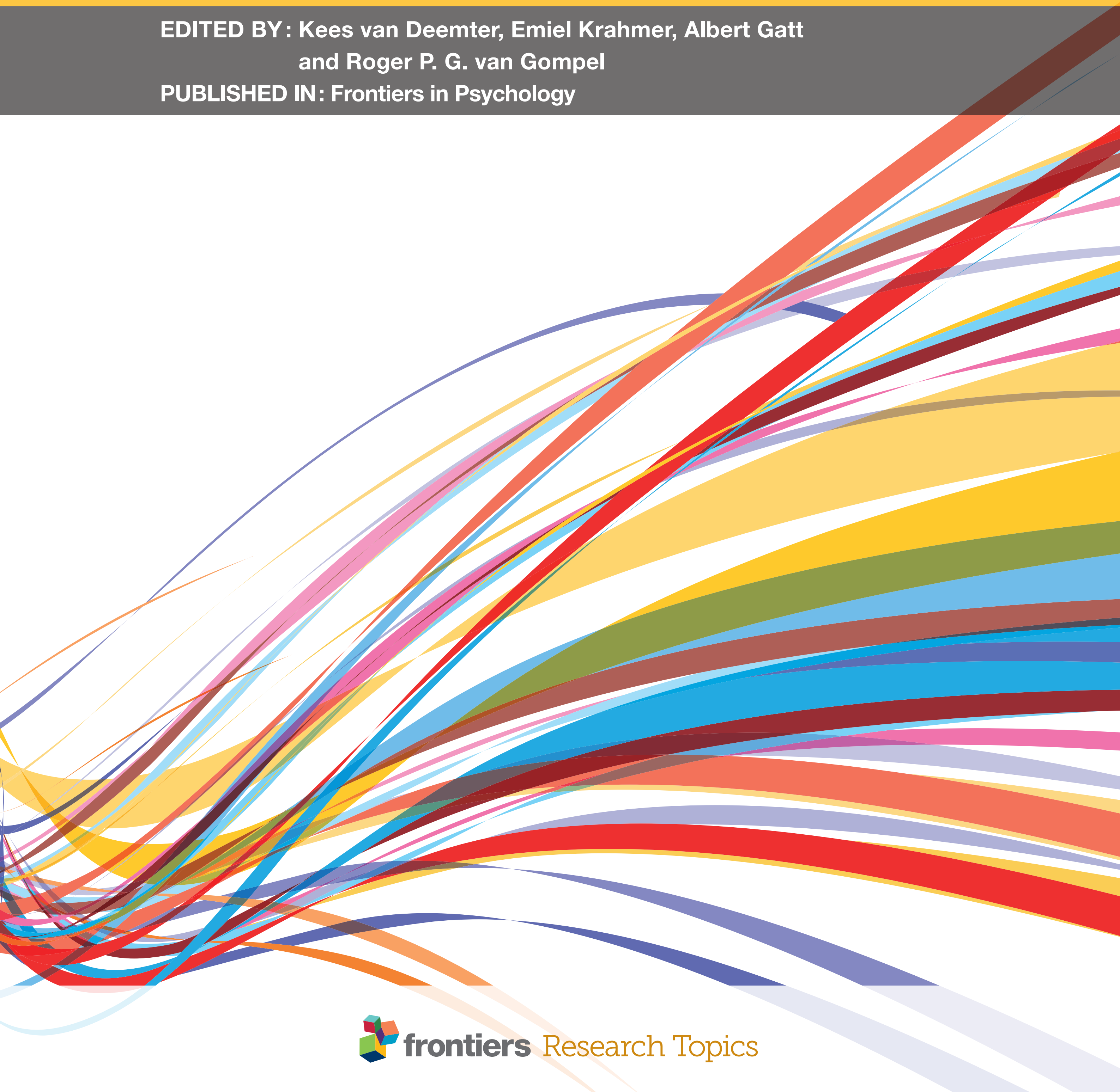
MODELS OF REFERENCE

EDITED BY: Kees van Deemter, Emiel Krahmer, Albert Gatt
and Roger P. G. van Gompel

PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-160-9
DOI 10.3389/978-2-88945-160-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

MODELS OF REFERENCE

Topic Editors:
Kees van Deemter, University of Aberdeen, UK
Emiel Krahmer, Tilburg University, Netherlands
Albert Gatt, University of Malta, Malta
Roger P. G. van Gompel, University of Dundee, UK

To communicate, speakers need to make it clear what they are talking about. Referring expressions play a crucial part in achieving this, by anchoring utterances to things. Examples of referring expressions include noun phrases such as “this phenomenon,” “it,” and “the phenomenon to which this Topic is devoted.” Reference is studied throughout the Cognitive Sciences (from philosophy and logic to neuro-psychology, computer science and linguistics), because it is thought to lie at the core of all of communication.

Recent years have seen a new wave of work on models of referring, as witnessed by a number of recent research projects, books, and journal Special Issues. The Research Topic “Models of Reference” in Frontiers in Psychology is a new milestone, focussing on contributions from Psycholinguistics and Computational Linguistics. The articles in it are concerned with such issues as audience design, overspecification, visual perception, and variation between speakers.

Citation: van Deemter, K., Krahmer, E., Gatt, A., van Gompel, R. P. G., eds. (2017). Models of Reference. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-160-9

Table of Contents

SECTION 1. EDITORIAL

SECTION 2. AUDIENCE DESIGN & THEORY OF MIND USE

SECTION 3. OVERSPECIFICATION

SECTION 4. VISUAL CONTEXT AND INDIVIDUAL VARIATION

frontiers in Psychology

EDITORIAL
published: 05 December 2016
doi: 10.3389/fpsyg.2016.01855

Editorial: Models of Reference

Kees van Deemter[1]*, Emiel Krahmer[2], Albert Gatt[2,3] and Roger P. G. van Gompel[4]

[1] Department of Computing Science, University of Aberdeen, Aberdeen, UK, [2] Tilburg Centre for Cognition and Communication, Tilburg University, Tilburg, Netherlands, [3] Institute of Linguistics, University of Malta, Msida, Malta, [4] School of Social Sciences, University of Dundee, Dundee, UK

Keywords: psycholinguistics, language production, referring expressions, computational models, audience design

Editorial on Research Topic

Models of Reference

OPEN ACCESS

Edited and reviewed by:
*Manuel Carreiras,
Basque Center on Cognition, Brain
and Language, Spain*

****Correspondence:***
*Kees van Deemter
k.vdeemter@abdn.ac.uk*

Specialty section:
*This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology*

Received: *03 November 2016*
Accepted: *09 November 2016*
Published: *05 December 2016*

Citation:
*van Deemter K, Krahmer E, Gatt A
and van Gompel RPG (2016) Editorial:
Models of Reference.
Front. Psychol. 7:1855.
doi: 10.3389/fpsyg.2016.01855*

1. INTRODUCTION

To communicate, speakers need to make it clear what they are talking about. *Referring expressions* play a crucial part in achieving this, by anchoring utterances to things. Examples of referring expressions include noun phrases such as "this phenomenon," "it," and "the phenomenon to which this Topic is devoted." Reference is studied throughout the Cognitive Sciences (van Deemter, 2016).

Recent years have seen a new wave of work in this area, as witnessed by a number of journal Special Issues.[1]The Research Topic "Models of Reference" in *Frontiers in Psychology* is a new milestone, focussing on contributions from Psycholinguistics and Computational Linguistics.

Unsurprisingly given the journal, the response to our Call for Papers has focussed predominantly on psycholinguistic work. A majority of submissions dealt with language production, as opposed to comprehension. In what follows, we summarize the papers accepted for this Research Topic, stressing some of the main themes emerging, including audience design (Section 2); overspecification (Section 3); visual perception, and variation between speakers (Section 4). We end with some general observations.

2. AUDIENCE DESIGN AND THEORY OF MIND USE

Successful communication requires that speakers and hearers take each other's knowledge into account, yet recent studies have questioned the extent to which they are able to do this (e.g., Horton and Keysar, 1996; Keysar et al., 2003). The present Research Topic shows that reference is still a key battleground in this debate.

Ibarra and Tanenhaus, for instance, investigate to what extent interlocutors are able to switch between different ways to conceptualize an object, as a function of the conversational setting in which the dialogue takes place; for example, a part of an object may be called a "wrench" in one setting (because it looks like one) but a "leg" in another (because that's its function). The authors conclude that switching between conceptualizations takes place with remarkable ease: "conceptual pacts are fluid temporary agreements." Branigan et al. focus on 8- to 10-year-old children, investigating the extent to which these are able to assess Common Ground developed through prior linguistic context, and whether this is sensitive to variations in prior interactions with the listener. Similar to adults, children adjust the length of their referring expressions depending on

[1] For instance, in *Language, Cognition, and Neuroscience* in 2014 (http://www.tandfonline.com/toc/plcp21/29/8), and in *Topics in Cognitive Science* in 2012 (http://onlinelibrary.wiley.com/doi/10.1111/tops.2012.4.issue-2/issuetoc). See also Gatt et al. (2014).

whether, in the preceding conversation, their conversational partner was a side-participant (who both saw the object and heard the expression), an overhearer (who only heard the expression), or a new participant. Nadig et al. put the spotlight on adults with Autism Spectrum Disorder. Although these were less likely than neurotypical speakers to adapt their referential behavior to their interlocutor, what stands out from their work, is how subtle the differences in behavior were, given that referential Audience Design may be thought to be particularly challenging for people with autism (see also earlier studies such as Begeer et al., 2010).

A new strand of work seeks to model interlocutors' reasoning about Common Ground more precisely than before. Our Topic contains two examples, which complement each other neatly: The article by Gegg-Harrison and Tanenhaus follows on from Heller et al. (2012), asking what an interlocutor can figure out, from earlier dialogue moves, about the likelihood that her interlocutor is familiar with a given proper name. Kutlák, et al. focus on situations where the hearer does not know the name of the referent, so it is crucial that suitable properties are selected for inclusion in the referring expression. The authors offer a computational model of a speaker's assessment of the likelihood that the hearer knows about any given property of a referent (e.g., that Darwin wrote "On the Origin of Species").

Horton and Brennan, finally, step back to discuss the information stored by interlocutors about what has been said over the course of a dialogue, asking what it is that they remember of it, and whether these meta-representations are subject to specific constraints. Proposing a synthesis of their earlier work, they argue that "any representations that capture information about others' perspectives are likely to be relatively simple and subject to the same kinds of constraints on attention and memory that influence other kinds of cognitive representations." The use of generic psychological mechanisms is a theme that can be discerned in other papers as well, as we will see.

3. OVERSPECIFICATION

Much research has focussed on speakers' inclination to include more information in referring expressions than hearers require for identifying the referent (Pechmann, 1989; Engelhardt et al., 2006), a phenomenon known as overspecification. Overspecification features strongly in the contributions by Rubio-Fernández, by Tarenskeen et al., and by Westerbeek et al.

Westerbeek et al. examine an idea that, though it has been examined in the past (e.g, Sedivy, 2003), has recently come to the fore, namely that properties that are *atypical* for a particular type of object are particularly likely to be included in a reference to the object. While these authors confirm earlier findings, focussing especially on color, they also find that the color typicality effect is moderated by color diagnosticity: it is strongest for high-color-diagnostic objects (i.e., objects with a simple shape). Atypicality is likewise discussed by Rubio-Fernández. Scrutinizing the well-attested propensity of color terms to be used in overspecification, she found that this propensity is modulated by factors such as typicality and the extent to which color can facilitate object recognition. Tarenskeen et al. focus their take on these issues on another dimension of variation between properties, namely whether they express an absolute property (such as color) or a relative one (such as size).

The contribution of Brodbeck et al. shows, following on from earlier studies such as Engelhardt et al. (2011), how brain studies that measure ERP can track the time course of the process whereby a hearer comprehends a referring expression. Among other things, their work suggests that when we read an overspecified expression, then even after we have identified the referent, we *reactivate* the corresponding representation when processing additional words. The authors argue that this might explain the benefits that overspecification (cf. Section 3) has been shown to have in some situations.

Finally, while the paper by Pogue et al. is concerned with overspecification, it is also relevant to our proposed theme of rationality. These authors asked how listeners might make rational use of linguistic information despite the fact that the linguistic input to which they are exposed often includes more, or less, information than what is necessary and sufficient for a given referential intention. Their model suggests that part of the answer lies in hearers' ability to adapt their expectations to a particular speaker. This brings us to a final strand of work discussed in this Topic.

4. VISION AND INDIVIDUAL VARIATION

4.1. Visual Perception and Salience

Reference, of course, is not tied to any particular perceptual modality—we routinely talk about things we have never seen. Yet much of our knowledge of the production of referring expressions has focussed on visual domains. A number of articles in this Topic extend this body of knowledge, with an emphasis on links with visual perception.

The papers by Clarke et al. and by Baltaretu et al. exemplify this line of work, inquiring how the perceptual configuration of a visual display influences reference. Baltaretu et al. find that when referring to an object using a spatial relation (e.g., "the ball between the doll and the train"), speakers' choice of relatum depends in part on its spatial location in the scene. Clarke et al. focus on scenes with visual clutter. Their main finding is that the visual salience of objects affects their order of mention in a description, a finding that is mirrored by an experiment in which salient objects are shown to be detected faster if they are mentioned earlier. Taken together, these findings support the view that visual perception is tightly coupled with language use (Tanenhaus et al., 1995).

4.2. Individual Variation

Variation between people is a basic observation in psychology, and studies of language production are starting to focus on this reality. Using machine learning, Kibrik et al. offer a model of the choice between different types of referring expression. They focus directly on the issue of variation, examining its implications for computational models of language production. The contribution by Hendriks takes a more theoretical approach, hypothesizing that differences in cognitive capacity can explain an important part of the observed variation between speakers. Hendriks discusses a model based on the cognitive architecture ACT-R (e.g., Anderson, 1993), which focusses on individual differences in processing speed and working memory capacity, arguing that these factors can be predictive of both underspecification and overspecification, and of listeners's tendency to misunderstand referring expressions as well. The contribution by Peters et al., finally, argues that although pronouns and repeated references are processed in different ways, these differences can be explained by general memory principles such as interference, suppression, and competition. This idea is consonant with those of Horton and Brennan (see above), who emphasize generic psychological mechanisms as well.

5. CONCLUSION

Collectively, the papers in this Research Topic show that the study of reference is continuing to attract a large amount of interesting work. Speaking in general, we were struck by an openness to new research methods and paradigms, including neuro-cognitive methods and computational modeling.

Focussing more specifically on the aforementioned themes, we continue to see a large amount of work on **audience design**, but rather than investigating whether adults are cooperative or not (as in most previous research), researchers now realize that it is not an all-or-nothing issue and investigate what information speakers use for being cooperative, they study different participant populations and build models of cooperative behavior in a range of different communicative situations. As for **overspecification**, where earlier work has tended to single out particular properties (e.g., color) as having a high propensity for being used in overspecification, the papers in this Topic paint a subtler picture, where a property may be overspecification-prone in some situations but not in others. Research on **individual variation**, finally, is still in its infancy, but the paper by Hendriks shows one promising direction in which this research may go, by focusing on general memory principles and known cognitive differences between individuals. We expect that these issues will be fleshed out in future by new computational models as well as by brain studies.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The first author acknowledges support from the EPSRC for the project RefNet (An interdisciplinary Network Focussing on Reference), EP/J019615/1.

REFERENCES

Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Begeer, S., Malle, B., Nieuwland, M. S., and Keysar, B. (2010). Using theory of mind to represent and take part in social interactions: comparing individuals with high-functioning autism and typically developing controls. *Eur. J. Dev. Psychol.* 7, 104–122. doi: 10.1080/17405620903024263

Engelhardt, P., Bailey, K. G., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Engelhardt, P. E., Demiral, S. B., and Ferreira, F. (2011). Over-specified referring expressions impair comprehension: an ERP study. *Brain Cogn.* 77, 304–314. doi: 10.1016/j.bandc.2011.07.004

Gatt, A., Krahmer, E., Van Deemter, K., and Van Gompel, R. (2014). Models and empirical data for the production of referring expressions. *Lang. Cogn. Neurosci.* 29, 899–911. doi: 10.1080/23273798.2014.933242

Heller, D., Skovbroten, K., and Tanenhaus, M. (2012). To name or to describe: shared knowledge affects referential form. *Top. Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01182.x

Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117. doi: 10.1016/0010-0277(96)81418-1

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89, 25–41. doi: 10.1016/S0010-0277(03)00064-7

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 98–110. Incremental speech production and referential overspecification.

Sedivy, J. (2003). Pragmatic versus form–based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguist. Res.* 32, 3–23. doi: 10.1023/A:1021928914454

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. G. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

van Deemter, K. (2016). *Computational Models of Referring: A Study in Cognitive Science*. Cambridge, MA: MIT Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 February 2016
doi: 10.3389/fpsyg.2016.00213

Do You Know What I Know? The Impact of Participant Role in Children's Referential Communication

Holly P. Branigan[1]*, *Jenny Bell*[1] *and Janet F. McLean*[2]

[1] *Department of Psychology, The University of Edinburgh, Edinburgh, UK,* [2] *Division of Psychology, Abertay University, Dundee, UK*

For successful language use, interlocutors must be able to accurately assess their shared knowledge ("common ground"). Such knowledge can be accumulated through linguistic and non-linguistic context, but the same context can be associated with different patterns of knowledge, depending on the interlocutor's participant role (Wilkes-Gibbs and Clark, 1992). Although there is substantial evidence that children's ability to model partners' knowledge develops gradually, most such evidence focuses on non-linguistic context. We investigated the extent to which 8- to 10-year-old children can assess common ground developed through prior linguistic context, and whether this is sensitive to variations in participant role. Children repeatedly described tangram figures to another child, and then described the same figures to a third child who had been a side-participant, an overhearer, or absent during the initial conversation. Children showed evidence of partner modeling, producing shorter referential expressions with repeated mention to the same partner. Moreover, they demonstrated sensitivity to differences in common ground with the third child based on participant role on some but not all measures (e.g., description length, but not definiteness). Our results suggest that by ten, children make distinctions about common ground accumulated through prior linguistic context but do not yet consistently deploy this knowledge in an adult-like way.

Keywords: children, dialogue, common ground, referential communication, participant role

Edited by:
Roger P. G. Van Gompel,
University of Dundee, UK

Reviewed by:
Aparna Nadig,
McGill University, Canada
Elizabeth Nilsen,
University of Waterloo, Canada
Ruth Margaret Kempson,
King's College London, UK

****Correspondence:***
Holly P. Branigan
holly.branigan@ed.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *31 July 2015*
Accepted: *03 February 2016*
Published: *23 February 2016*

Citation:
Branigan HP, Bell J and McLean JF (2016) Do You Know What I Know? The Impact of Participant Role in Children's Referential Communication. Front. Psychol. 7:213. doi: 10.3389/fpsyg.2016.00213

INTRODUCTION

Learning to use language successfully requires more than simply acquiring words to express particular concepts and the grammar to combine those words to form particular propositions; it also involves learning when to use which words and which grammatical forms to particular listeners so that the speaker's meaning is appropriately communicated to the addressee. Adults appear to use information from a range of sources to shape the way in which they design their utterances to be easily understood. Research with children suggests that they begin to show sensitivity to a conversational partner's perspective in their language use from an early age, but it is still unclear what factors they take into account when modeling their partner's knowledge, and exactly how such beliefs about their partner's knowledge are manifested in their language production. In this research, we consider whether 8 to 10-year-old children are able to draw appropriate inferences about their partners' knowledge on the basis of their partners' participation in previous dialogue, and examine how such inferences might be reflected in the language they produce.

Speakers can refer to things in many different ways; for example, the same entity can be described as *a dog* or *the fluffy Labrador from down the road.* This is particularly the case for entities with low codability such as tangrams, which can usually be conceptualized in very different ways (e.g., as a skater vs. a chicken) depending on a speaker's perspective (Clark and Wilkes-Gibbs, 1986). How then does a speaker choose a particular referring expression to use? Substantial research has suggested that speakers' choices involve *audience design* (Bell, 1984), or a consideration of what the addressee is likely to understand. To do this, speakers draw on their *common ground,* the knowledge that they believe themselves to share with their listeners.

Clark and Marshall (1981) identified three possible sources of shared knowledge. One important source is beliefs about the cultural communities to which their listeners belong (Fussell and Krauss, 1992). For example, if the speaker believes that she and the addressee are both members of the University of Edinburgh community, she can assume that they share knowledge about particular buildings, people, procedures, and so on. Adults consistently use such beliefs to choose between alternative referring expressions (e.g., whether to refer to a building as "*McEwan Hall*" vs. "*The round building with the dome*"; Isaacs and Clark, 1987).

But assumptions about shared knowledge can also be based on evidence that is tied to particular interactions. Speakers can make reference to the physical context in which they and their listeners are situated, and assume that an object (or indeed any kind of experience) that is physically co-present, and of which listeners might be aware, constitutes part of their common ground. Similarly, they can make reference to previous physical co-presence (e.g., common past experiences).

More relevantly for our concerns, they can also make reference to preceding linguistic context, in other words the language that the speaker and listener have previously used together (in the current or previous conversations), and the meanings that they have jointly established for these utterances. Thus when a speaker produces an utterance in the presence of a particular listener (e.g., "*This tangram looks like a chicken*"), its linguistic content (e.g., words, syntax, phonology) becomes part of their linguistic common ground. In addition, their shared understanding of the meaning of this utterance (the *situation model* that it maps onto; Zwaan and Radvansky, 1998) becomes part of their linguistic common ground.

However, Clark and Wilkes-Gibbs (1986) suggested that the reference of an utterance (i.e., the link between the linguistic expression and the particular referent to which it is intended to refer) becomes part of common ground only following a collaborative process that requires the participation of both speaker and addressee to establish a mutual belief that the addressee has correctly understood the speaker's intended reference. Only when the speaker and addressee mutually accept that the addressee has understood the speaker sufficiently can the reference enter their common ground. Once this mutual acceptance has been reached, the speaker can subsequently assume that the addressee will understand that reference correctly if she uses it again. The speaker and addressee therefore form a *referential pact* for how to refer to the object (e.g., as a chicken).

Accordingly, Clark and Wilkes-Gibbs (1986) showed that when speakers (*Directors*) described a set of tangrams to the same partners (*Matchers*), they initially tended to produce extended descriptions and indefinite references (e.g., "*looks like a person who's ice skating, except they're sticking two arms out in front*"), which were shaped by feedback from their partners over a number of turns (just 18% of initial descriptions were immediately accepted by the Matcher, in what Clark and Wilkes-Gibbs termed a *basic exchange*) until both participants were satisfied that understanding had been achieved. When they subsequently referred to the same tangrams, speakers tended to use definite and considerably shorter references (e.g., "*the ice skater*"), and addressees were able to accept these without requiring further elaboration. Brennan and Clark (1996) showed that speakers also produced fewer *hedge* expressions (indicating provisionality; e.g., *sort of, a bit*) on repeated reference. The result of these adaptations was that communication became faster and more efficient, requiring fewer words and fewer turns.

These findings suggest that speakers' choice of referring expressions was affected by their previous discourse with a partner (see also Garrod and Anderson, 1987). Brennan and Clark (1996) subsequently showed that these effects were partner-specific: Speakers used the same referring expressions repeatedly with the same partner, even when the context made them over-informative. Referential pacts also affect comprehension, with addressees showing slower reaction times to identify referents when the speaker violates a referential pact by using a new term for a referent, even if it is otherwise an appropriate description (e.g., Metzing and Brennan, 2003; Shintel and Keysar, 2007; Brown-Schmidt, 2009a).

Clark and Carlson (1982) noted that dialogues may also involve roles other than speaker and addressee. For example, a person may be a ratified participant in a conversation, but not be directly addressed by the speaker. Clark (1992) proposed that such *side participants* accrue common ground in the same way as speakers and addressees; they share responsibility for tracking what is said and for ensuring that they understand the speaker. The speaker can therefore assume that anything that forms part of their common ground with an addressee also forms part of their common ground with a side participant. In contrast, *overhearers* are not ratified participants in the conversation: Although they hear what the speaker says, they are not under any responsibility to maintain a record of the discourse or to ensure that they have understood the speaker (and by corollary, do not have privileges to collaborate to reach understanding). They do not therefore accumulate common ground with the speaker in the same way as the addressee, and the speaker cannot assume that overhearers have access to the same common ground as an addressee. In accord with this proposal, Schober and Clark (1989) showed that overhearers had a poorer understanding of a director's descriptions in a tangram task than addressees, even when they heard the entire dialogue and were given the advantage of being able to pause and replay the director's descriptions, suggesting that they did not have access to the same common ground as addressees.

Wilkes-Gibbs and Clark (1992) showed that such differences in common ground associated with different participant roles were reflected in speakers' referential behavior. Speakers repeatedly described a set of tangrams to a partner (as in Clark and Wilkes-Gibbs, 1986; Matcher A), before describing the same set to a different partner (Matcher B), who had previously played one of four roles: a silent side participant (seated next to the Director during her interactions with Matcher A), an omniscient bystander (watching and listening on a monitor in a separate room), an overhearer (seated behind the Director in such a way that they could hear the director and matcher A's conversation but could not see any of the referents), or a naïve participant (seated outside the experimental room engaged in a separate task, and so unable to see or hear any of the conversation).

On the first round following the changeover, Directors were fastest and used fewest words with former side participants, followed by omniscient bystanders; they were slowest and used most words with overhearers and naïve participants. They also produced significantly more indefinite references (and correspondingly fewer definite references) when Matcher B had been a simple bystander or naïve participant than a side participant or omniscient bystander. These results are consistent with Directors making different assumptions about the common ground that they shared with Matcher B on the basis of participant role. When Matcher B had been a side participant or omniscient bystander, Directors treated them similarly to Matcher A. In contrast, Directors treated overhearers in a similar manner to naïve participants, assuming little or no common ground. Thus, although overhearers had been able to hear descriptions, Directors acted as if this information was insufficient for successful reference without knowledge of the referent that each description was anchored to.

In sum, there is evidence that adult speakers are sensitive to variations in the information that they share with their addressees, and assume different levels of common ground depending on their addressee's participant role in previous discourse. Although there may be some leakages (e.g., failures to initially accommodate common ground during the earliest stages of processing; Horton and Keysar, 1996; Lane and Ferreira, 2008), adults tend to produce referential expressions that reflect these assumptions, with respect to the semantic content of their referring expressions (e.g., use of alternative conceptualizations), the amount of information they provide (e.g., shorter vs. longer referring expressions), and the form in which they express this information (e.g., use of definite vs. indefinite referring expressions).

Does children's referential communication similarly reflect their beliefs about what their partner is likely to understand? Certainly, children appear to be aware from an early age that people may have different knowledge from their own (e.g., Perner et al., 1987; Astington and Gopnik, 1991), and reflect this in their non-verbal communicative behavior (e.g., pointing and gesturing; Perner et al., 1987; Liszkowski et al., 2008). But is this awareness reflected in their language use, and what kinds of evidence are their beliefs about shared and unshared knowledge based on?

Some studies have shown that children, like adults, adapt their language production to reflect beliefs about their addressees' likely knowledge based on community membership. For example, children younger than five adapt the grammar and vocabulary of their utterances depending on their addressee's identity (e.g., producing less complex grammar and vocabulary when addressing a baby or a child than an adult; Shatz and Gelman, 1973; Sachs and Devin, 1976; Hansson et al., 2000; Hoff, 2010). This is consistent with a coarse degree of audience design that does not require detailed modeling of an addressee's knowledge, but can be based on broad distinctions (e.g., Galati and Brennan, 2010).

Children also show sensitivity to common ground based on past and present physical co-presence, though their ability to accommodate this information in their referential communication varies. Many studies have suggested that children are poor at producing unambiguous referential expressions to pick out one object from a complex array of objects with similar characteristics until well into school age (e.g., Glucksberg et al., 1966; Krauss and Glucksberg, 1969; Dickson, 1982; Deutsch and Pechmann, 1982; Lloyd et al., 1995, 1998). For example, Deutsch and Pechmann (1982) found that half of 6-year-olds (and a fifth of 9-year-olds) were unable to produce unambiguous referring expressions on their first attempt (e.g., saying the *"red one"* in a context involving several red objects), although they were responsive to their addressees' feedback.

Equally, Anderson et al. (1991) found that 7- to 8-year-olds (and 9- to 10-year-olds to a less marked degree) in route-giving dialogues that involved mismatching maps tended to inappropriately introduce new referents using definite references. Thus younger children presupposed that referents were shared with their addressees, rather than collaboratively establishing their shared status and a referential pact for how to refer to them (and their addressees were equally poor at providing feedback when referents were not in fact shared).

Such difficulties have been interpreted in terms of egocentric processing (Piaget, 1959). However, they may also reflect children's difficulties in determining relevant dimensions of contrast (e.g., Sonnenschein and Whitehurst, 1984). Recent studies have shown that by five, children can produce referring expressions whose content reflects the information that the child believes to be in perceptual common ground when the context makes it easier for the child to discriminate privileged from mutually shared knowledge. Hence 5-year-olds are more likely to produce an adjective to unambiguously pick out an object when there is a competitor object visible to both the child and their addressee than when the competitor is visible only to the child (e.g., Nadig and Sedivy, 2002; Bahtiyar and Küntay, 2009; Nilsen et al., 2009). Matthews et al. (2006) found that 3- and 4-year-olds also adapted the form of their referring expressions, such that their choice of (more informative) lexical NPs (e.g., "*The clown*") vs. (less informative) pronouns (e.g., *"he"*) to refer to an entity was affected by whether the referent was visible to the addressee or not, although they still frequently failed to do so (e.g., 4-year-olds inappropriately produced pronouns on a third of trials where the referent was visually inaccessible).

In addition, older children adapt their referential behavior on the basis of previous physical co-presence, suggesting that in this age group the ability to engage in audience design based on shared physical context is not contingent on the context being concurrently available for consultation. Sonnenschein (1988) found that 6- to 9-year-old children produced referential expressions that contained more (possibly redundant) information when pretending to describe a toy for a stranger or friend with no shared experience than for a friend with whom they shared a common experience. Taken together, these results suggest that by school age, children are able to assess common ground based on past and current physical co-presence to at least some extent. Moreover, these assessments may affect both the amount of information provided in, and the form of, children's referential expressions, although they may not do so consistently and children's referential expressions may not always be optimal (e.g., in terms of redundancy).

There has been much less research on children's assumptions about common ground based on linguistic co-presence, and the extent to which these constrain referential processing. Unlike common ground based on concurrent physical co-presence, where the relevant context is available for consultation, common ground based on linguistic co-presence requires the child to be able to maintain and continuously update relevant information in memory. As such, it might be both more complex and more effortful to track. In comprehension, Matthews et al. (2010) found that 3- and 5-year-olds were slower to pick up and move an object when their partner referred to it using a different name (e.g., "*truck*") than she had previously used to refer to it (e.g., "*car*"), than when a different partner, who had not previously named the object, referred to it using the new name (cf. Metzing and Brennan, 2003; Shintel and Keysar, 2007; Brown-Schmidt, 2009a). Graham et al. (2014) replicated these effects when the referential pact violation related to use of an adjective (e.g., "fluffy dog" vs. "spotted dog," for a dog that was both fluffy and spotted), rather than different conceptualizations of the object at a categorical level. These results suggest that in comprehension, even pre-school children are sensitive to linguistic common ground, and specifically the referential pacts that they and a particular partner have established in previous discourse.

However, although these results suggest that children track the linguistic common ground that they have established with a partner, and are able to use this information to constrain comprehension by the age of four, children do not appear to use linguistic common ground to guide their production of referring expressions until later in development. Köymen et al. (2014) had 4- and 6-year-old children describe objects to a partner, and then describe the same objects in a different visual context to the same or a different partner. Six-year-olds were sensitive to whether or not they had previously established relevant referential pacts with a partner: If they had, they re-used the referring expression they had previously (tacitly) agreed; if they had not, they chose the referring expression that was most appropriate given the context of the array. Thus, their referential choices reflected audience design based on linguistic common ground. In contrast, 4-year-olds consistently produced referring expressions that were appropriate given the visual context, and showed no sensitivity to whether or not they and their addressee had previously established relevant linguistic common ground.

Other evidence suggests that children's ability to accumulate and use linguistic common ground appropriately continues to develop over a prolonged timecourse. Studies involving tasks in which children must communicate interactively about complex domains (e.g., maps with mismatching landmarks, mazes that involve complex spatial arrays) show that school-aged children experience difficulties in accurately modeling their partner's knowledge and responding to feedback up to the age of 11 and beyond (Anderson et al., 1991, 1994; Garrod and Clark, 1993). Garrod and Clark (1993) found that pairs of 7- to 8-year-olds sometimes converged on the same referring expressions, but without the same reference (e.g., both using *where you/I started*, but to refer to different locations), suggesting that their choice of referential expression was not based on a representation of common ground that included the crucial connection between a referring expression and its referent. Moreover, Anderson et al. (1994) found that even at the age of 13, a substantial minority of children performed no better than 7- to 8-year-olds. Clearly, children's ability to accumulate and flexibly exploit common ground when they speak in dialogue does not reach full maturity for many years.

Overall, the evidence suggests that, like adults, children maintain a representation of the language that they have previously used with a particular conversational partner, and that this model of linguistic common ground affects their referential processing to at least some extent from a young age, although the ability to use this information appropriately during the production of referential expressions appears to continue to develop into the teen years. But at what age do children develop a mature understanding of the accumulation of linguistic common ground? In particular, when do they become sensitive to participant roles, and understand that people accumulate common ground differently based on their participant role within a dialogue? All previous research has focused on how children use common ground accumulated within a dyadic dialogue involving just a speaker and an addressee. Although this research casts light on children's assumptions about common ground between speakers and addressees, it is not informative about children's awareness of the more general relationship between participant roles and the establishment of shared knowledge, in other words that listeners might develop shared knowledge with the speaker differently depending on whether they are licensed participants in the conversation or not.

The data from dyadic dialogues is compatible with children having an adult-like understanding that when a speaker proposes something and the addressee accepts it, the speaker's proposal becomes part of the linguistic common ground of all participants. But it also compatible with children having an impoverished understanding of the accumulation of linguistic common ground based on a simple distinction between having been the addressee of a particular utterance or not, or alternatively on having been present when something was said or not. In the former case, children might wrongly assume that someone who had previously been a side-participant would not have access to the language that was used in that conversation (or its interpretation); in the latter case, children might understand that an addressee who was not previously present would not have access to the language that was used in that conversation (and its interpretation), but they might wrongly assume that an overhearer who had been present during that conversation would also have access to it.

To distinguish these alternatives, we carried out an experiment in which eight- to ten-year old children played a tangram-description and -matching task with a partner, as in Wilkes-Gibbs and Clark (1992). One child was designated the Director, and played the game with another child (Matcher A) over four rounds (A1–4); the same Director then played the same game, using the same tangrams, with a second child (Matcher B; rounds B1-4). We manipulated Matcher B's participant role during the first four rounds, in order to vary the linguistic common ground shared by the Director and Matcher B during their subsequent interaction.

In the *side-participant* condition, Matcher B was seated next to the Director (and had the same view of the Director's tangrams as the Director) throughout the Director's rounds with Matcher A. Thus, Matcher B was able to hear all the references made and also verify whether these references were successfully resolved (through Matcher A's responses and the final outcome of each round). In the *overhearer* condition, Matcher B was seated in the same room but approximately 2 meters behind the Director with her back to the Director and Matcher; she could therefore hear references and exchanges with Matcher A, but could not see either player's tangrams, hence which tangram was being referred to. In the *naïve participant* condition, Matcher B was seated outside the experimental room; the Director and Matcher B therefore shared no common ground. Following Wilkes-Gibbs and Clark (1992), our primary measures were the total time taken for Directors and Matchers to match the set of tangrams each round, the number of correctly placed tangrams (measures of collaborative communicative success), and—in order to assess Directors' initial audience design based on their *a priori* beliefs about the Matcher's knowledge—the mean number of words per tangram that Directors used in their initial utterances before they received any formative feedback from the Matcher. As additional measures, we examined Directors' use of definite or indefinite reference (an index of whether Directors believed reference to be shared) and number of hedges (an index of their commitment to a particular conceptualization for a referent) in their initial utterances, as well as the number of basic exchanges (where the Director described a tangram and the Matcher immediately accepted this description; an index of the adequacy of the Director's audience design from the Matcher's perspective, i.e., whether the Matcher found the Director's initial description sufficient to identify the tangram).

Given previous findings that school-aged children are sensitive to the accumulation of linguistic common ground with an addressee, we expected that rounds A1–4 would show the same pattern as found in adults (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986), specifically a tendency toward greater efficiency and shorter, definite descriptions that the addressee immediately accepts, which is assumed to reflect the exploitation of common ground accumulated with the partner over the course of the interaction.

However, our main interest is Directors' behavior in round B1, when interacting with a new partner. If children have an adult-like expectation that all participants within a dialogue (whether silent or actively involved) assume responsibility for their part in the collaborative process, then we would expect the same pattern as Wilkes-Gibbs and Clark (1992) found in adults. When playing with a former side-participant, the Director should assume that Matcher B has accumulated as much common ground during rounds A1–4 as both the Director and Matcher A. She should therefore assume that Matcher B has access to the referential pacts that she established with Matcher A, and so tend to use shorter, definite references with few hedges, and her descriptions should tend to be immediately comprehensible to Matcher B (yielding the same pattern of basic exchanges as with Matcher A). We would therefore expect the Director and Matcher B to take a similar amount of time and to have a similar level of accuracy as the Director and Matcher A did on round A4.

When playing the game with a former overhearer, the Director should assume that although she may have heard the linguistic expressions that were used, she would not have grounded their reference, and therefore does not have access to the referential pacts that she had established with Matcher A. She should therefore treat overhearers in the same way as naïve participants (as in Wilkes-Gibbs and Clark, 1992), yielding longer and more informative descriptions than on round A4, with more indefinite references and more hedges. Because she has not yet established a referential pact with Matcher B, we might expect that Matcher B would be less likely to find her initial description comprehensible, resulting in slower times, fewer basic exchanges, and lower accuracy than in round A4.

If however children have a non-adult-like understanding of the way in which linguistic common ground is accumulated, then we would expect a different pattern. If they make a simple distinction based on having been the addressee of a particular utterance or not, then in all three conditions they should treat their addressees as if they had no access to linguistic common ground, using longer and more informative descriptions, with fewer definite descriptions and more hedges, than round A4.

If instead children make a simple distinction based on having been present when something was said or not, then they should treat side-participants and overhearers (both of whom were present during rounds A1–4) differently from naïve participants (who were not). In that case, Directors should produce similar descriptions in the side-participant and overhearer conditions as on round A4, but in the naïve participant condition they should produce longer and more informative indefinite descriptions with more hedges. Because Directors would be erroneously overestimating addressees' knowledge in the overhearer condition, we might expect that accuracy in this condition would be reduced compared to A4 (and total time might be increased).

These predictions are based on the assumption that children's beliefs about linguistic common ground will be manifested in the same ways as in adults. However, the literature reviewed above shows that children may sometimes show audience design with respect to some aspects of language (e.g., use of lexical NPs vs. pronouns) but not others (use of definite vs. indefinite NPs). It may therefore be the case that children will show effects on some measures but not on others. Such a pattern would be informative about the extent to which children and adults manifest common ground in their linguistic behavior in the same way.

METHODS

Participants

Seventy-two children aged between 8 and 10 years (mean: 9 years 7 months) recruited from a junior school in Nottinghamshire, UK, participated in the experiment (i.e., 8 groups per condition). This study was approved by the University of Edinburgh Psychology Department Ethics committee. Parents provided informed written consent for children's participation, and children provided verbal consent.

Materials

The experimental items were eight tangrams taken from Wilkes-Gibbs and Clark (1992). Each tangram was printed in black on cream card (15 by 20 cm) and was laminated. Nine copies were made of the remaining eight tangrams to form the experimental sets; one copy was used by the matcher, and eight copies were used by the director (one set for each of the four rounds with each matcher). The tangrams in each of the director's sets were placed in numbered envelopes in a randomized order. Two further tangrams were used for demonstration and practice purposes.

To engage children with the task, we also provided a cardboard pyramid with a "jungle adventurer" theme; if children correctly matched four or more tangrams in a round, they could move an adventurer figure up a level on the pyramid. To ensure that children acting as Matcher B in the overhearer condition remained focused, and to give them a defined role (so that they were not perceived as an eavesdropper), we also prepared a handout featuring four pyramids, each with eight levels, for Matcher B to color in when they thought the Director and Matcher A had matched one tangram. We also prepared a booklet containing three mazes and games with a "jungle adventurer" theme, to occupy children who were not currently engaged in the game (Matcher B in the naïve participant condition for rounds A1–4; Matcher A in all three conditions for rounds B1-B4).

Design

The experiment used a 3 × 2 mixed design, with Participant Role (side participant, overhearer or naïve participant) as a between-subjects factor and Round (A1 and A4, or A4 and B1) as a within-subjects factor.

Procedure

Groups of three children were taken into the experimental room and told that they would play a "jungle adventurer" game, in which they would match ancient symbols to crack a secret code and reveal hidden treasure. Groups were randomly allocated to one of the experimental conditions. The children drew lots to decide roles. The Director and Matcher A took their seats, and Matcher B sat next to the experimenter where she could observe the table. A table divider in the middle of the table prevented Director and Matcher from seeing each other's cards. The children were told that they would play the game in two stages. First, the Director would describe the symbols in each envelope to Matcher A, so that the Matcher could put them in the same order; they could talk as much as needed to match the figures quickly and accurately. The Director and Matcher A would do this for four envelopes, all of which included the same symbols but in a different order. The Director would then do the same with Matcher B for a further four rounds.

One tangram was used as an example to familiarize the children with the figures; a second was used as a practice, to ensure that the Director provided sufficiently detailed descriptions. After the practice, Matcher B was informed of his role (in earshot of the director) before being taken to his corresponding position as a side participant, overhearer or naïve participant. Side participants were told that they would be able to see and hear what the Director was doing in the game, although they would not be playing it yet themselves. Overhearers were told that they would not be able to see anything but that they would be able to hear; they were also given the task of monitoring the Director and Matcher A's progress by coloring in levels on the pyramid sheet. Naïve participants were told that they would not be able to hear or see anything as they would be completing the activity booklet outside the room.

The Director opened the first envelope and laid out the cards in order. The Director and Matcher A then began their four rounds. After each round, the experimenter checked the accuracy of the card positions, and provided feedback about how many were correctly placed. After the Director and Matcher A had completed their four rounds (A1–4), Matcher B took the place of Matcher A (and in all conditions Matcher A took the overhearer's seat and was given the activity booklet to complete).

The children's interactions were audio-recorded using a tape recorder. The experiment took approximately 45 min to complete.

Scoring

All rounds with Matcher A and Matcher B were timed from start to finish, using a stopwatch. Success was measured at the end of each round, by counting how many tangrams the children had correctly matched and converting this to a percentage.

Rounds A1, A4, and B1 were transcribed by the second author, and were independently coded by two coders who were ignorant of the experimental hypotheses (Cohen's kappa, a measure of inter-coder reliability, is reported below; in all cases, there was very high agreement). Disagreements were resolved by discussion. The dependent variables were based on the director's initial descriptions, before they received any feedback from the matcher. Feedback was classified as any sort of interruption or interaction (e.g., a question or contribution) that led to modification by the director; backchannel responses (e.g., *yeah*) that encouraged to the director to continue were not classified as feedback. Given that the initial description could only have been influenced by the director's *a priori* beliefs about their matcher's level of knowledge, this was judged to provide a more accurate and uncontaminated measure of audience design based on assumptions about linguistic common ground.

We recorded the *mean number of words* that the director used to introduce each figure before feedback from the matcher was recorded. Following Wilkes-Gibbs and Clark (1992), we coded directors' initial references as *definite reference* if they included the form *the x*, *this/that x*, or *the one with x,* or no article at all (e.g., *the next one is x*), and as *indefinite reference* if they included the form *a/an x.* (Other references were descriptive, e.g., *it has an X*; Cohen's kappa = 1). We further measured the number of *hedges* that directors produced, focusing on four specific forms: *"sort of," "kind of," "a bit,"* and *"-ish"* (Cohen's kappa = 0.985). We note that children also produced very high numbers of another type of hedge, namely *like* (e.g., *the next one's got erm two like half triangles*). Although these hedges are potentially highly informative, many examples could not be reliably discriminated as hedges vs. expressions of similarity (e.g., *it's got like a leg*), and we therefore did not code their use. Finally, we recorded the number of *basic exchanges* between directors and matchers. An exchange was coded as a basic exchange if the matcher immediately accepted the director's initial description without refashioning it in any way, so that the director immediately continued to the next tangram (Clark and Wilkes-Gibbs, 1986; Cohen's kappa = 0.991). We give examples of each coding category below.

1a. Definite reference: *The seal*
1b. Indefinite reference: *A man sat with no arms and no legs, like he's sat down.*
1c. Hedge [underlined]: *Like a head <u>kind of</u> triangle thing*
1d. Basic exchange: Director: *Zombie*
Matcher: *Yeah*

RESULTS

We analyzed seven dependent variables: Mean total time (in seconds) per round; mean number of tangrams successfully identified (out of eight) per round; mean number of words per tangram in the Director's initial description per round; frequency of a definite referring expression in the Director's initial description per round; frequency of an indefinite referring expression in the Director's initial description per round; frequency of a hedge expression in the Director's initial description per round; and frequency of a basic exchange per round. Twenty-five data points (i.e., references to tangrams) were excluded because the Director did not refer to the relevant tangram (all involved the final tangram in a round, where the correct tangram could be identified by elimination).

We used mixed effects models to analyze the data. When the dependent variable was continuous, we modeled the response using linear mixed effects models, and when the dependent variable was binomial (basic exchange vs. not a basic exchange), we modeled the responses using logit mixed effects models (Baayen et al., 2008; Jaeger, 2008). For each binomial model, we were interested in predicting the probability of a positive response in the different conditions (i.e., that the children used a basic exchange). For all analyses there were two fixed effects (Participant Role and Round). Participant Role had three levels (naïve participant vs. side participant vs. overhearer), and Helmert coding was used to explore how the presence of Matcher B affected the Director's referring behavior. The first contrast compared the naïve participant condition, where Matcher B was not present during rounds A1–A4, to the mean of the overhearer and side participant conditions, where Matcher B was present during rounds A1–A4. A second contrast compared the overhearer and side participant conditions. Round had two levels for each analysis (A1 vs. A4 and A4 vs. B1); deviation coding was used to contrast each level. Full random effects models would not converge, so Round was removed from the random effects structure. Only significant (or marginal; $p < 0.1$) results are reported.

To confirm whether children showed the same patterns as found in previous research on adults when repeatedly describing the same referents to the same partner (e.g., Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986; Wilkes-Gibbs and Clark, 1992), we began by comparing rounds A1 and A4. Note that in these analyses, differences between Participant Role conditions would reflect incidental differences (e.g., in individual Directors' communicative skills), because the Participant Role manipulation was irrelevant at this stage. Any such differences in rounds A1–3 would moreover be irrelevant to our key questions, which hinge on differences between the final round with Matcher A (i.e., A4) and the first round with Matcher B.

However, our primary interest was in Directors' different assumptions about linguistic common ground with Matcher B as a function of Matcher B's previous participant role during rounds A1–4 with Matcher A. Hence the main comparisons of interest are those examining Directors' changes in behavior between their final round with Matcher A (A4) and their first round with Matcher B (B1).

Total Time Taken Per Round (Table 1)

Rounds A1–A4

The model comparing Rounds A1 and A4 revealed a significant main effect of Round ($\beta = -178.13$, $SE = 23.6$, $t = -7.54$, $p_z < 0.001$[1]): Round A4 was completed faster than A1 (For this and all other linear mixed model analyses, p-values were calculated using a normal approximation). There was a marginal interaction between Participant Role and Round when naïve participant was contrasted with the two other conditions ($\beta = 92.25$, $SE = 50.1$, $t = 1.84$, $p_z = 0.07$): There was a greater reduction between A1 and A4 in the naïve participant condition (240 s) than the mean of the other two conditions (147 s). However, a model that included simple main effects for only round A4 showed that there was no difference between Participant Role conditions in round A4 (both $p_z > 0.30$).

A second set of analyses examined whether there was a reduction in time across rounds A1–A4 in each Participant Role condition. For these analyses, rounds A1, A2, A3, and A4 were included in a model and Round was coded using polynomial coding. There was a significant linear trend for each Participant Role, with A1 being the slowest round and A4 being the fastest (all $p_z < 0.01$).

[1] For this and all other linear mixed model analyses, p-values were calculated using a normal approximation.

TABLE 1 | Mean total time taken (sec) and percentage of tangrams correctly matched, by Round and Participant Role; standard deviation is in square brackets.

	Naïve Participant			Overhearer			Side-Participant		
	A1	**A4**	**B1**	**A1**	**A4**	**B1**	**A1**	**A4**	**B1**
Total time (sec)	349 [91.7]	109 [57.7]	315 [190.4]	293 [166.5]	118 [48.4]	219 [165.8]	256 [115.7]	137 [78.6]	115 [87.0]
% correct	70.3 [30.6]	84.4 [18.6]	75.0 [22.2]	56.3 [32.7]	64.1 [27.1]	53.1 [25.7]	42.2 [28.3]	70.3 [29.1]	82.8 [20.0]

Rounds A4-B1

The model comparing rounds A4 and B1 with Helmert contrasts between the naïve participant condition and the other two conditions, and between the overhearer vs. side participant conditions, revealed a significant main effect of Round ($\beta = 95.17$, $SE = 27.0$, $t = 3.53$, $p_z < 0.001$); overall, round B1 was completed more slowly than A4. There was also a significant interaction between Participant Role and Round when naïve participant was contrasted with the other two conditions ($\beta = -165.7$, $SE = 57.3$, $t = -2.89$, $p_z < 0.01$), and a marginal interaction for overhearer vs. side participant ($\beta = -122.63$, $SE = 66.1$, $t = 1.85$, $p_z = 0.06$). Round B1 was 206 and 101 s slower than A4 when the matcher was a former naïve participant or overhearer respectively, but 22 s faster than A4 when the matcher was a former side participant.

A model that included simple main effects for round B1 showed a significant difference between the naive participant and the other two conditions ($\beta = -145.3$, $SE = 53.2$, $t = -2.73$, $p_z < 0.01$); and a marginal difference between overhearer and side participant ($\beta = 101.73$, $SE = 54.8$, $t = 1.86$, $p_z = 0.06$). Round B1 was slower when Directors were describing to a naïve participant than when they were describing to an overhearer or side participant. Directors were also slower when Matcher B had been an overhearer than when they had been a side participant.

Number of Tangrams Correctly Matched (Table 1)

Rounds A1–A4

The model comparing accuracy on rounds A1 and A4 revealed a significant main effect of Participant Role for naïve participant vs. the other two conditions ($\beta = -1.50$, $SE = 0.66$, $t = -2.27$, $p_z < 0.05$); there were more correctly matched tangrams in the naïve participant condition than in the other two conditions. There were no differences between the overhearer and side participant conditions. There was also a main effect of Round ($\beta = 1.33$, $SE = 0.54$, $t = -2.49$, $p_z < 0.05$; matchers correctly matched more tangrams in round A4 than A1. A model that included simple main effects for round A4 showed a marginal difference between the naïve participant and other two conditions ($\beta = -1.41$, $SE = 0.76$, $t = -1.87$, $p_z = 0.06$), with more correct tangrams in the naïve participant condition.

Rounds A4-B1

The model comparing Rounds A4 and B1 including Helmert contrasts showed a marginal main effect of Participant Role for naïve participant vs. the other two conditions ($\beta = -1.01$, $SE = 0.57$, $t = -1.78$, $p_z = 0.07$); participants correctly matched more tangrams in the naïve participant condition than the overhearer and side participant conditions (irrespective of round). A model analysing simple main effects for round B1 only showed a significant difference for the overhearer vs. side participant conditions ($\beta = -2.19$, $SE = 1.03$, $t = -2.13$, $p_z < 0.05$), with more correct tangrams in the side participant condition.

Mean Number of Words Per Tangram (Table 2)

Rounds A1–A4

There was a significant main effect of Round ($\beta = -11.83$, $SE = 1.29$, $t = -9.13$, $p_z < 0.001$); Directors produced fewer words in their initial descriptions (prior to feedback) in A4 than in A1. There was a significant two-way interaction between Participant Role and Round when naïve participant was contrasted with the two other conditions ($\beta = 8.39$, $SE = 2.75$, $t = 3.05$, $p_z < 0.01$); Directors' initial descriptions reduced more from A1 to A4 in the naïve participant condition. There was also a significant two-way interaction between Participant Role and Round when overhearer was contrasted with side participant ($\beta = -6.52$, $SE = 3.17$, $t = -2.05$, $p_z < 0.05$); Directors' initial descriptions reduced more from A1 to A4 in the overhearer condition.

However a model that included simple main effects for round A4 showed no differences between the naïve participant and other two conditions, nor between overhearer and side participant (all $p_z > 0.48$); by A4, Directors in all conditions were producing a similar number of words to describe the tangrams in their initial descriptions.

Rounds A4-B1

The model comparing rounds A4 and B1 revealed a significant main effect of Round ($\beta = -6.70$, $SE = 1.23$, $t = -5.45$, $p_z < 0.001$); overall, Directors produced more words in their initial descriptions on their first round with matcher B than their last round with matcher A. There were also significant interactions between Participant Role and Round in the naïve participant condition contrasted with the other two conditions ($\beta = 15.89$, $SE = 2.61$, $t = 6.08$, $p_z < 0.001$), and for overhearer vs. side participant ($\beta = -6.07$, $SE = 3.01$, $t = -2.02$, $p_z < 0.05$). Directors used more words in their initial descriptions when addressing a new partner (Matcher B) who had been a naïve participant, and to a lesser extent when addressing a former overhearer. In contrast, Directors in the side participant condition used fewer words in their initial descriptions when describing tangrams to Matcher B for the first time than when describing the same tangrams to Matcher A for the fourth time.

TABLE 2 | Mean number of words per tangram in Director's initial description, by Round and Participant Role; standard deviation is in square brackets.

	Naïve Participant			Overhearer			Side-Participant		
	A1	A4	B1	A1	A4	B1	A1	A4	B1
Words/tangram	32.40 [19.3]	15.13 [12.4]	33.05 [23.4]	26.94 [17.6]	14.17 [7.4]	18.67 [10.1]	23.48 [16.6]	17.63 [14.2]	15.91 [11.7]

A model analysing simple main effects for round B1 showed a significant difference only between the naïve participant condition contrasted with the other two conditions (β = −14.49, *SE* = 4.30, *t* = −3.37, $p_z < 0.001$). Directors produced more words when they knew that Matcher B was a naïve participant.

Definite and Indefinite References (Table 3)

Rounds A1–A4

Definite references

As no directors produced definite references in round A1, these data were not suitable for logit mixed effect models. All three Participant Roles showed an increase in definite references between A1 and A4. A model analysing simple main effects for round A4 showed no difference between participant role conditions ($p_z > 0.96$).

Indefinite references

There was a significant main effect of Round (β = −0.70, *SE* = 0.29, *t* = −2.39, $p_z < 0.05$); children produced fewer indefinite references on A4 than A1. There were also two-way interactions between Participant Role and Round when the naïve participant condition was contrasted with the other two conditions (β = −1.28, *SE* = 0.59, *t* = −2.19, $p_z < 0.05$), and for overhearer vs. side participant (β = 3.13, *SE* = 0.75, *t* = −4.18, $p_z < 0.001$); Directors with a side participant initially produced the highest number of indefinite references but then substantially reduced their indefinite references between A1 and A4. Directors in the naïve participant condition produced the lowest number of indefinite references in A1, and both they and Directors in the overhearer condition showed little change across rounds. A model analysing simple main effects on round A4 showed only a marginal difference between the overhearer and side participant conditions ($p_z = 0.08$), with more indefinite references in the overhearer than side participant condition (36 vs. 19).

Rounds A4-B1

Definite references

The model comparing rounds A4 and B1 revealed a significant main effect of Round (β = 4.23, *SE* = 1.51, *Z* = 2.81, $p_z < 0.001$); Directors produced fewer definite references on their first round with matcher B than their last round with matcher A. The interaction with Participant Role was not significant, despite the greater number of definite references produced in B1 by Directors in the side participant condition. Closer inspection revealed that all definite references in the side participant condition (across all rounds) were produced by the same three directors.

Indefinite references

The model comparing rounds A4 and B1 showed a significant main effect of Round (β = −0.70, *SE* = 0.30, *Z* = 2.31, $p_z <$ 0.05); Directors produced more indefinite references in round B1 than in A4, irrespective of participant role.

Hedges (Table 4)

Rounds A1–A4

The model revealed a significant main effect of Round (β = −0.23, *SE* = 0.05, *t* = −4.91, $p_z < 0.001$), indicating that the children produced fewer hedges on the fourth round with matcher A. There was also a significant interaction between Round and Participant Role for overhearer vs. side participant (β = −031, *SE* = 0.11, *t* = −2.71, $p_z < 0.01$): Directors reduced their number of hedges between Rounds A1 and A4 to a greater extent in the overhearer condition (23 vs. 3).

Rounds A4-B1

The model comparing rounds A4 and B1 revealed a significant main effect of Round (β = −0.11, *SE* = 0.04, *t* = −2.94, $p_z < 0.01$); Directors produced more hedges on B1 than on A4, irrespective of participant role.

Basic Exchanges (Table 3)

Rounds A1–A4

The model revealed a significant main effect of Round (β = 1.17, *SE* = 0.28, *Z* = 4.14, $p_z < 0.001$); although there was a high proportion of basic exchanges even in round A1 (at least half of all descriptions in every condition), this number increased from A1 to A4. A model analysing simple main effects on round A4 showed no difference between participant role conditions on round A4 ($ps_z > 0.19$).

Rounds A4-B1

The model comparing rounds A4 and B1 revealed a significant interaction between Participant Role and Round for the contrast between the overhearer and side participant conditions (β = 2.23, *SE* = 0.76, *Z* = −2.92, $p_z < 0.01$); basic exchanges decreased when the director changed partners in the overhearer condition, but increased in the side participant condition. A model analysing simple main effects for round B1 showed a marginal difference only between the naïve participant condition contrasted with the other two conditions (β = 1.15, *SE* = 0.63, *Z* = 1.84, $p_z = 0.06$). Directors produced more words when they knew that Matcher B was a naïve participant.

GENERAL DISCUSSION

Our experiment set out to examine 8-10-year-old children's assumptions about the accrual of linguistic common ground,

TABLE 3 | Frequency of definite and indefinite references in Director's initial description, and basic exchanges, by Round and Participant Role; percentage of total tangrams per round is in parentheses, and standard deviation is in square brackets.

	Naïve Participant			Overhearer			Side-Participant		
	A1	**A4**	**B1**	**A1**	**A4**	**B1**	**A1**	**A4**	**B1**
Def. refs	0 (0.0%) [0.0]	20 (31.3%) [3.2]	3 (4.7%) [1.1]	0 (0.0%) [0.0]	15 (23.4%) [2.8]	5 (7.8%) [2.7]	0 (0.0%) [0.0]	24 (37.5%) [4.1]	22 (34.4%) [3.8]
Indef. refs	23 (35.9%) [2.5]	24 (37.5%) [3.3]	30 (46.9%) [3.1]	33 (51.6%) [2.7]	36 (56.3%) [3.4]	40 (62.5%) [3.0]	42 (65.6%) [3.4]	19 (29.7%) [3.0]	25 (39.1%) [3.0]
Basic exch.	32 (50.0%) [1.6]	50 (78.1%) [1.9]	41 (64.1%) [2.6]	44 (68.8%) [2.4]	52 (81.3%) [2.0]	43 (67.2%) [2.7]	42 (65.6%) [2.9]	46 (71.9%) [2.7]	53 (82.8%) [1.6]

TABLE 4 | Frequency of hedges in Director's initial description, by Round and Participant Role; standard deviation is in square brackets.

	Naïve Participant			Overhearer			Side-Participant		
	A1	**A4**	**B1**	**A1**	**A4**	**B1**	**A1**	**A4**	**B1**
Hedges	19 [3.9]	1 [0.4]	8 [2.1]	30 [5.1]	7 [1.5]	15 [3.0]	6 [2.1]	3 [0.7]	8 [2.4]

and how this would affect their language use in a referential communication task. Specifically, we were interested in whether they would display an adult-like appreciation of differences in how common ground accumulates based on distinctions in listeners' participant roles. Previous research has focused exclusively on speaker-addressee pairings, and has suggested that children make assumptions that addressees have access to shared linguistic information in a way that people who have been absent from the conversation do not. However, such evidence does not demonstrate that children have a mature understanding of how differences in participant roles and the responsibilities associated with being a licensed participant in a conversation affect the accrual of common ground. Children might instead use simpler distinctions when assessing common ground, either overestimating its accumulation (by assuming that all listeners have access to it, irrespective of whether they are licensed participants), or underestimating its accumulation (by assuming that only addressees have access to it). We tested these possibilities by having children play a game in which they described the same set of tangrams repeatedly to another child and then described them again to a third child who had seen and heard the initial conversation, had only heard the conversation, or had neither seen nor heard the conversation.

We first consider the results from rounds A1 to A4 (before any change in partner), and their implications for children's accumulation of common ground in speaker-addressee pairings. The fact that Directors produced progressively shorter descriptions for the tangrams as they repeatedly described them to the same Matcher is consistent with previous research on adults (e.g., Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986). This shortening occurred for Directors' initial descriptions, prior to receiving any formative feedback from the Matcher, and therefore suggests that Directors were exploiting their knowledge of the linguistic common ground that they had built up with the Matcher to design, *a priori,* referring expressions that the Matcher would be able to understand. Analyses of Matchers' tangram-matching accuracy and time taken to complete each round suggest that Directors effectively exploited common ground in this way: Despite the shortening in initial referring expressions across rounds (from 27.6 words per tangram in A1 to 15.65 words in A4), tangram-matching accuracy increased (from 56.3 to 72.9%), and the time taken to complete each round decreased (from 299 to 121 s). Additionally, Matchers were more likely to accept the shorter initial descriptions in round A4 immediately without requiring further information, than the longer initial descriptions in A1 (77.1 vs. 61.5% basic exchanges respectively).

In other words, with increasing interaction with their partners, Directors produced shorter descriptions that were nevertheless communicatively more effective and more efficient. Our results therefore show that when 8–10-year-olds encounter novel objects with no conventional label, they are able not only to initially generate appropriate referring expressions for them in collaboration with their addressee, but also to subsequently draw on this shared linguistic information to design more concise but comprehensible references. These findings therefore extend previous research showing that children make use of referential pacts when referring to objects with conventional labels (Köymen et al., 2014).

Children also showed sensitivity to linguistic common ground in other aspects of their language. Their use of definite references (presupposing shared knowledge) changed across rounds. In round A1, where the Director and Matcher had no linguistic common ground, Directors never used definite references; in A4, where they had accrued common ground over the preceding three rounds, they used definite references on 30.7% trials. Children also produced fewer references that included expressions of uncertainty (hedges such as *sort of*) as they developed common ground with their partner, dropping from 18 hedges per round in A1 to 4 hedges per round in A4. Overall, then, the results of rounds A1–A4, prior to the manipulation of prior participant role, demonstrate that when interacting with a single partner, children of this age are able to track and use linguistic common ground, for at least some aspects of their language, in ways that enhance communication.

Unexpectedly, there were some differences between conditions in rounds A1–A4 (e.g., more correctly matched tangrams in the Naïve Participant condition), even though at this point the role of Matcher A was equivalent across conditions. It seems most likely that such differences reflect coincidental variations in individual children's performance rather than any effect of the experimental manipulation. As is clear from our results, and consistent with previous findings (e.g., Anderson et al., 1994), there were substantial individual differences in children's performance (e.g., in A1, the mean number of words in Directors' initial descriptions ranged from 7 to 55, and Matchers' tangram accuracy ranged from 0 to 100%).

We now turn to our main question of interest, namely children's assumptions about the accumulation of common ground based on differences in participant role. Our critical analyses therefore concerned changes in Directors' behavior between rounds A4 (their last round with Matcher A) and B1 (their first round with Matcher B). We were interested in whether children would show an adultlike pattern of treating former side participants in B1 in the same way as they had treated Matcher A in A4 (i.e., assuming equal access to common ground), in contrast to former overhearers and naïve participants; or would show a non-adultlike pattern, either treating all new partners alike (assuming no access to common ground, and yielding uniform differences between behavior on A4 and B1, irrespective of participant role), or treating both side participants and overhearers on B1 in the same way as they had treated Matcher A on A4 (with only naïve participants being treated differently). Our results suggest that although children have some understanding that linguistic common ground accumulates differently according to distinctions in listeners' participant roles, their understanding is not yet fully adultlike. They also suggest (in conjunction with analyses of rounds A1–A4) that the ways in which children draw on linguistic common ground in their language use differs from adults.

The primary evidence that children are sensitive to differences in participant role comes from analyses of the length of Directors' initial descriptions. These suggest that Directors made a tripartite distinction between the information available to former side participants, naïve participants, and overhearers. When addressing former side participants for the first time, they produced initial referring expressions that were very similar (and in fact, slightly shorter) than those that they produced when addressing Matcher A for the fourth time (A4: 17.63 words; B: 15.91 words). This is consistent with Directors assuming that side participants in a dialogue had access to the same linguistic common ground as addressees, and so could benefit from the same kind of concise referring expression.

Their ability to produce appropriate referring expressions in B1 on the basis of linguistic common ground accrued over rounds A1–A4 is supported by the fact that total time taken to complete the round did not increase when they first interacted with a new partner (indeed, it decreased by 22 s from A4 to B1) and at the same time tangram matching accuracy did not decrease (rather, increased by 12.5%); additionally, the number of turns in which Matchers were able to accept the initial description immediately did not decrease (rather, increased by 9.9%). Hence, Directors behaved as though they had the same common ground with former side participants as with former addressees; moreover, their ensuing referring expressions were communicatively effective, showing successful audience design on the basis of these assumptions.

In contrast, when Directors addressed former naïve participants for the first time, they produced initial referring expressions that were considerably longer than those that they had produced when addressing Matcher A for the fourth time (A4: 15.13 vs. B1: 33.05 words; this difference was significantly larger than the side participant/overhearer conditions). This result suggests that when Directors designed their referring expressions in B1, they assumed—prior to receiving any feedback from the Matcher—that a new Matcher who had been outside the room during the initial rounds required more information than Matcher A had required in A4; in other words, they assumed that naïve participants did not have access to the same common ground as addressees. Accordingly, the total time taken to complete the round increased by 206 s from A4 to B1, though tangram matching accuracy did not differ.

Directors also appeared to make a further distinction concerning the information available to former overhearers. Their initial referring expressions for Matcher B in round B1 were slightly longer than those for Matcher A in round A4 (A4: 14.17; B1: 18.67). The significant difference in the mean number of words per tangram in A4 vs. B1 in the overhearer and side participant conditions implies that Directors did not assume that former overhearers and former side participants had access to the same common ground. However, nor did they appear to treat former overhearers as having the same (lack of) knowledge as naive participants. The fact that Directors only slightly increased the length of their initial referring expressions suggests that they overestimated former overhearers' knowledge, and that this impacted negatively on communication. The total time taken to complete the round increased by 101 s, but more critically tangram accuracy in B1 was lower than in the side-participant condition; additionally, Matchers were less able to immediately accept Directors' initial referring expressions—indicating perceived understanding—in the overhearer condition than in the side participant condition (note that the two conditions did not differ on either measure in A4). It appears that Directors did not fully grasp the limited extent to which prior exposure to referring expressions alone, without simultaneous exposure to the reference of those expressions, was likely to facilitate subsequent comprehension.

In sum, evidence from the length of Directors' initial referring expressions suggests that children made largely accurate assumptions about the extent to which former naïve participants and former side participants had access to linguistic common ground, and designed their referring expressions accordingly, but also provides some suggestion that they were less accurate in gauging former overhearers' shared knowledge, with an apparent tendency to overestimate it. This pattern differs from that found in adults, who tend to treat addressees who previously had access to the linguistic content, but not the reference, of a prior dialogue in the same way as addressees who had no previous access to a prior dialogue (Wilkes-Gibbs and Clark, 1992).

However, this sensitivity to participant role is not borne out in other aspects of our data. Overall, Directors did not behave differently to former side participants, overhearers and naïve participants with respect to their use of definite and indefinite referring expressions, or hedges. Based on previous research on adults, we might have expected that the use of definites (implying shared knowledge) would decrease from A4 to B1 in the naïve participant condition relative to the side participant and overhearer conditions, and that the use of indefinites would conversely increase (Wilkes-Gibbs and Clark, 1992). We might also have expected the use of hedges (indicating provisionality prior to agreement on a referential pact) to increase (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996).

Although children did show differences on all three measures when interacting for the first time with a new partner (a lower use of definites, and a higher use of indefinites and hedges, in B1 than A4), these differences were uniform across conditions when the groups were considered as a whole. (However, we note that there was considerable variation between individual children in the use of hedges and definite references, suggesting that these aspects of language use might be particularly subject to individual differences in development; see also (Anderson et al., 1994; Nilsen et al., 2009), for further evidence of individual differences in dialogue skills). When considered alongside the evidence discussed above that children are nevertheless sensitive to differences in the accrual of linguistic common ground according to participant role, this pattern suggests that children do not accommodate these differences in their language in the same way as adults. In this study, assumptions about common ground manifested consistently in the length of children's referring expressions, but not the form of those expressions.

The conclusion that children and adults linguistically manifest common ground differently is supported by evidence from rounds A1 to A4. Although Directors showed increased use of definite expressions between rounds A1 and A4, definites still formed less than a third of their references in A4, and their use of indefinite expressions did not decrease between A1 and A4, remaining around two fifths of all references (41.6 vs. 41.2%). In contrast, Wilkes-Gibbs and Clark (1992) found that by the sixth round, Directors used definite references on 86.5% of trials and indefinite references on 4.2%. Thus, in adults' dialogue, there was a very strong relationship between the accrual of common ground and the use of definite expressions, whereas in our children this tendency was considerably weaker.

These results are consistent with other evidence suggesting that children do not use definite and indefinite reference in the same way as adults (e.g., Maratsos, 1974; Warden, 1976; Anderson et al., 1991). Most such research has found that children tend to overuse definite expressions, for example when first mentioning a referent that is unknown to the addressee. In these studies, children tend to incorrectly assume that their addressee has access to the same set of referents as themselves. Our study suggests that in a different context, where children knew that they had access to the same set of referents as their addressee but these referents did not have conventional names (and so could be conceptualized in multiple ways), children tended to underuse definite expressions, relative to adults. That is, they did not tend to use definite references that depended on (non-conventional) referential pacts (e.g., referring to *the rabbit*), although the shortening of referential expressions with increased common ground suggests that they were aware of, and exploited, these pacts (e.g., referring to a tangram in terms of its similarity to a rabbit). These results suggest that even at the age of 8–10 years, children's use of definite and indefinite referring expressions may differ from that of adults.

Finally, we consider other aspects of our results that suggest further disparities between children's and adults' referential processing in dialogue, focusing on rounds A1–A4 (to exclude any influences associated with changes in partner and participant role). Children's performance was consistently and considerably poorer than adults. Children's error rates ranged from 42.7% (A1) to 27.1% (A4). In contrast, Clark and Wilkes-Gibbs (1986) found error rates of around 2% in their studies (with a larger item set, which should have increased the likelihood of misidentification).

The high error rate is not surprising in itself, but it is indicative of the children's limited ability to detect and/or resolve misunderstandings. For an error to occur, Directors and Matchers must have terminated the process of presenting and accepting a reference inappropriately: The Director must have failed to detect that the Matcher had selected the wrong tangram, and the Matcher must have failed to realize that the Director was referring to a tangram other than the one they had selected. That is, they both inaccurately believed that the Matcher had understood the Director correctly, and therefore allowed the dialogue to move on. Thus although Matchers' increasingly accurate and faster performance across rounds in response to progressively shorter initial descriptions demonstrates that the Director and Matcher were able to build up and exploit common ground to some extent, the relatively high error rate overall indicates that this ability was still immature and far from adultlike.

This conclusion is supported by evidence from the occurrence of basic exchanges in rounds A1–4. Clark and Wilkes-Gibbs (1986) found that with adult participants, basic exchanges occurred relatively infrequently in the first round of the task, where participants had to identify novel objects for which they had as yet established no common ground, but became highly frequent in later rounds once common ground had been established (round 1: 18%; round 4: 80%). Thus adult Directors and Matchers tended to be cautious in their assumptions of mutual understanding, and to initially exchange multiple turns to establish confidence that understanding had been successfully achieved. In contrast, our children showed very high levels of basic exchanges even in the very first round (A1: 61.5%). Clearly, in these trials Matchers believed that they had understood the Directors, and Directors believed that Matchers had understood them—the instructions, the structure of the game, and the feedback provided by the experimenter after each round all ensured that children were aware that the Matcher must identify and place in the appropriate position the specific tangram described by the Director—but the tangram accuracy data show that this belief was often incorrect.

These results are consistent with many previous findings suggesting that children may have difficulties both in evaluating their addressee's understanding and appropriately responding when acting as speaker, and in detecting their own failure to understand and/or appropriately requesting information when acting as addressee (e.g., Bearison and Levey, 1977; Ironsmith and Whitehurst, 1978; Whitehurst and Sonnenschein, 1981;

Anderson et al., 1991, 1994; Garrod and Clark, 1993; Lloyd et al., 1998). Thus in the same way that children may tend to overestimate the information that they share with a partner, they may also tend to overestimate the occurrence of mutual understanding.

Our study focused on one age group, and as such we cannot draw conclusions about the way in which, or age at which, children might come to develop adult-like behavior. Previous research suggests that even at the age of 13, a substantial minority of children continue to show behavior that differs from that found in experiments involving adults (Anderson et al., 1994). (Note, however, that most such experiments involve a relatively narrow population of highly educated individuals, i.e., college students, whose performance may not be representative of the adult population as a whole). The development of relevant dialogue skills may in part be dependent on the maturation of aspects of cognition such as executive function, such as the ability to inhibit one's own perspective. Certainly, research on both child and adult dialogue has implicated inhibitory control and working memory in online perspective-taking (Epley et al., 2004; Brown-Schmidt, 2009b; Nilsen et al., 2009; Lin et al., 2010).

However, in our experiment, the fact that Directors produced longer descriptions with naïve participants in B1 shows that they were able to suppress their own knowledge appropriately, suggesting that executive function (specifically inhibitory control) may be less relevant to our results, though working memory may have played some role. It therefore seems likely that the development of adult-like behavior cannot be reduced simply to the maturation of executive functions, and instead involves the development of a more elaborated understanding of what information is and is not shared by speakers and addressees on the basis of previous discourse (e.g., whether speaker and addressee share the reference of a referring expression).

We suggest that the interactions that children experience may play an important role in shaping this developmental process. Many studies have suggested that experience of communication breakdown and its subsequent resolution through formative feedback from listeners—whether at first-hand or through observation—may play an important role in improving young children's performance in dialogue tasks (e.g., Robinson and Robinson, 1981, 1985; Deutsch and Pechmann, 1982; Matthews et al., 2007, 2012). In principle, all of the interactions that children experience could give them valuable evidence about the accrual of common ground under different circumstances. However, given that formative feedback depends crucially on the listener, and given that school-aged children– as our and other studies show—are not always adept at gauging their own understanding and providing informative feedback, it may be the case that interactions with more mature language users (adults and near-adults) play a particularly important role in developing relevant skills and understanding even into the early teen years.

In conclusion, this research investigated what inferences 8–10-year-old children were able to draw about their partners' knowledge on the basis of their participation in previous dialogue. Our results suggest that by this age, children have some understanding that the accumulation of linguistic common ground is affected by participant role. In particular, they assume that side participants in a dialogue build up linguistic common ground (and have access to this common ground in subsequent dialogues involving the same speaker), and that overhearers do not have access to this information to the same extent. These assumptions are reflected in the amount of information that they provide in their referring expressions. However, our results also suggest that children are not fully adult-like at this age in both their understanding and their linguistic use of common ground. Children appear to overestimate the extent to which listeners who overhear but do not participate in a dialogue accumulate common ground, and do not use definiteness to reflect linguistic common ground in the same way as adults. These results, together with evidence of other limitations in children's dialogue skills (e.g., overestimations of mutual understanding) provide further evidence that learning to use language successfully in interaction is a slow process that continues to develop until well into the school years.

FUNDING

This research was supported by ESRC grant ES/F023359/1 awarded to JB and a British Academy/Leverhulme Trust Senior Research Fellowship awarded to HB.

ACKNOWLEDGMENTS

We thank all of the children, parents and teachers who participated in this study, and thank Kellie Hunter and Robin Macy for their assistance with data coding.

REFERENCES

Anderson, A. H., Clark, A., and Mullin, J. (1991). Introducing information in dialogues: forms of introduction chosen by young speakers and the responses elicited from young listeners. *J. Child Lang.* 18, 663–687.

Anderson, A. H., Clark, A, and Mullin, J. (1994). Interactive communication between children: learning how to make language work in dialogue. *J. Child Lang.* 21, 439–63.

Astington, J. W., and Gopnik, A. (1991). "Developing understanding of desire and intention," in *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, ed A. Whiten (Oxford: Basil Blackwell), 39–50.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Bahtiyar, S., and Küntay, A. C. (2009). Integration of communicative partner's visual perspective in patterns of referential requests. *J. Child Lang.* 36, 529–555. doi: 10.1017/S0305000908009094

Bearison, D. J., and Levey, L. M. (1977). Children's comprehension of referential communication: decoding ambiguous messages. *Child Dev.* 48, 716. doi: 10.2307/1128682

Bell, A. (1984). Language style as audience design. *Lang. Soc.* 13, 145–204. doi: 10.1017/S004740450001037X

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493.

Brown-Schmidt, S. (2009a). Partner-specific interpretation of maintained referential precedents during interactive dialog. *J. Mem. Lang.* 61, 171–190. doi: 10.1016/j.jml.2009.04.003

Brown-Schmidt, S. (2009b). The role of executive function in perspective taking during online language comprehension. *Psychon. Bull. Rev.* 16, 893–900. doi: 10.3758/PBR.16.5.893

Clark, H. H. (1992). *Arenas of Language Use. Arenas of Language Use*. Chicaco, IL: University of Chicago Press. Available online at: http://www.loc.gov/catdir/description/uchi051/92038874.html

Clark, H. H., and Carlson, T. B. (1982). Hearers and speech acts. *Language* 58, 332–373. doi: 10.1353/lan.1982.0042

Clark, H. H., and Marshall, C. R. (1981). "Definite reference and mutual knowledge," in *Elements of Discourse Understanding,* eds A. K. Joshi, B. L. Webber, and I. A. Sag (Cambridge: CUP), 10–63.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Deutsch, W., and Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition* 11, 159–184. doi: 10.1016/0010-0277(82)90024-5

Dickson, W. P. (1982). "Two decades of referential communication research: a review and meta-analysis," in *Verbal Processes in Children,* eds C. J. Brainerd and M. Pressley (New York: Springer), 1–33.

Epley, N., Morewedge, C. K., and Keysar, B. (2004). Perspective taking in children and adults: equivalent egocentrism but differential correction. *J. Exp. Soc. Psychol.* 40, 760–768. doi: 10.1016/j.jesp.2004.02.002

Fussell, S. R., and Krauss, R. M. (1992). Coordination of knowledge in communication: effects of speakers' assumptions about what others know. *J. Pers. Soc. Psychol.* 62, 378–391. doi: 10.1037/0022-3514.62.3.378

Galati, A., and Brennan, S. E. (2010). Attenuating information in spoken communication: for the speaker, or for the addressee? *J. Mem. Lang.* 62, 35–51. doi: 10.1016/j.jml.2009.09.002

Garrod, S., and Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition* 27, 181–218. doi: 10.1016/0010-0277(87)90018-7

Garrod, S., and Clark, A. (1993). The development of dialogue co-ordination skills in schoolchildren. *Lang. Cogn. Process.* 8, 37–41. doi: 10.1080/01690969308406950

Glucksberg, S., Krauss, R. M., and Weisberg, R. (1966). Referential communication in nursery school children: method and some preliminary findings. *J. Exp. Child Psychol.* 3, 333–342. doi: 10.1016/0022-0965(66)90077-4

Graham, S., a, Sedivy, J., and Khu, M. (2014). That's not what you said earlier: preschoolers expect partners to be referentially consistent. *J. Child Lang.* 41, 34–50. doi: 10.1017/S0305000912000530

Hansson, K., Nettelbladt, U., and Nilholm, C. (2000). Contextual influence on the language production of children with speech/language impairment. *Inter. J. Lang. Commun. Disord.* 35, 31–47. doi: 10.1080/136828200247232

Hoff, E. (2010). Context effects on young children's language use: The influence of conversational setting and partner. *First Lang.* 30, 461–472. doi: 10.1177/0142723710370525

Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117. doi: 10.1016/0010-0277(96)81418-1

Ironsmith, M., and Whitehurst, G. J. (1978). The development of listener abilities in communication: how children deal with ambiguous information. *Child Dev.* 49, 348–352

Isaacs, E. A., and Clark, H. H. (1987). References in conversation between experts and novices. *J. Exp. Psychol.* 116, 26–37. doi: 10.1037/0096-3445.116.1.26

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Köymen, B., Schmerse, D., Lieven, E., and Tomasello, M. (2014). Young children create partner-specific referential pacts with peers. *Development. Psychol.* 50, 2334–2342. doi: 10.1037/a0037837

Krauss, R. M., and Glucksberg, S. (1969). The development of communication: competence as a function of age. *Child Dev.* 40, 255–266.

Krauss, R. M., and Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychon. Sci.* 1, 113–114. doi: 10.3758/BF03342817

Lane, L. W., and Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: do cognitive demands override threats to referential success? *J. Exp. Psychol.* 34, 1466–1481. doi: 10.1037/a0013353

Lin, S., Keysar, B., and Epley, N. (2010). Reflexively mindblind: using theory of mind to interpret behavior requires effortful attention. *J. Exp. Soc. Psychol.* 46, 551–556. doi: 10.1016/j.jesp.2009.12.019

Liszkowski, U., Albrecht, K., Carpenter, M., and Tomasello, M. (2008). Infants' visual and auditory communication when a partner is or is not visually attending. *Infant Behav. Develop.* 31, 157–167. doi: 10.1016/j.infbeh.2007.10.011

Lloyd, P., Camaioni, L., and Ercolani, P. (1995). Assessing referential communication skills in primary school years: a comparative study. *Br. J. Dev. Psychol.* 13, 13–29.

Lloyd, P., Mann, S., and Peers, I. (1998). The growth of speaker and listener skills from five to eleven years. *First Lang.* 18, 81–103.

Maratsos, M. P. (1974). Preschool children's use of definite and indefinite articles. *Child Dev.* 45, 446–455. doi: 10.2307/1127967

Matthews, D., Butcher, J., Lieven, E., and Tomasello, M. (2012). Two- and four-year-olds learn to adapt referring expressions to context: effects of distracters and feedback on referential communication. *Top. Cogn. Sci.* 4, 184–210. doi: 10.1111/j.1756-8765.2012.01181.x

Matthews, D., Lieven, E., Theakston, A., and Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Appl. Psycholinguist.* 27, 403–422. doi: 10.1017/S0142716406060334

Matthews, D., Lieven, E., and Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: a training study. *Child Dev.* 78, 1744–1759. doi: 10.1111/j.1467-8624.2007.01098.x

Matthews, D., Lieven, E., and Tomasello, M. (2010). What's in a manner of speaking? Children's sensitivity to partner-specific referential precedents. *Dev. Psychol.* 46, 749–760. doi: 10.1037/a0019657

Metzing, C., and Brennan, S. E. (2003). When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *J. Mem. Lang.* 49, 201–213. doi: 10.1016/S0749-596X(03)00028-7

Nadig, A. S., and Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* 13, 329–336. doi: 10.1111/j.0956-7976.2002.00460.x

Nilsen, E. S., and Graham, S., a. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cogn. Psychol.* 58, 220–249. doi: 10.1016/j.cogpsych.2008.07.002

O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Develop.* 67, 659. doi:10.2307/1131839

Perner, J., Leekham, S., and Wimmer, H. (1987). Three-year olds' difficulty with false-belief: the case for a conceptual deficit. *Br. J. Dev. Psychol.* 5, 125–137.

Piaget, J. (1959). *Judgment and Reasoning in the Child.* Paterson: Littlefield, Adams and Co.

Robinson, E. J., and Robinson, W. P. (1981). Ways of reacting to communication failure in relation to the development of the child's understanding about verbal communication. *Eur. J. Soc. Psychol.* 11, 189–208. doi: 10.1002/ejsp.2420110206

Robinson, E. J., and Robinson, W. P. (1985). Teaching children about verbal referential communication. *Int. J. Behav. Dev.* 8, 285–299. doi: 10.1177/016502548500800304

Sachs, J., and Devin, J. (1976). Young children's use of age appropriate speech styles in social interaction and role-playing. *J. Child Lang.* 3, 81 98.

Schober, M. F., and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cogn. Psychol.* 21, 211–232. doi: 10.1016/0010-0285(89)90008-X

Shatz, M., and Gelman, R. (1973). The development of communication skills: modifications in the speech of young children as a function of listener. *Monogr. Soc. Res. Child Dev.* 38, 1–38.

Shintel, H., and Keysar, B. (2007). You said it before and you'll say it again: expectations of consistency in communication. *J. Exp. Psychol.* 33, 357–369. doi: 10.1037/0278-7393.33.2.357

Sonnenschein, S. (1988). The development of referential communication: speaking to different listeners. *Child Dev.* 59, 694–702.

Sonnenschein, S., and Whitehurst, G. J. (1984). Developing referential communication: a hierarchy of skills. *Child Dev.* 55, 1936–1945.

Warden, D. A. (1976). The influence of context on children's use of identifying expressions and references. *Br. J. Psychol.* 67, 101–112.

Whitehurst, G. J., and Sonnenschein, S. (1981). "The development of informative messages in referential communication: knowing when vs knowing how," in *Children's Communication Skills,* ed W. P. Dickson (New York: Academic Press), 127–141.

Wilkes-Gibbs, D., and Clark, H. H. (1992). Coordinating beliefs in conversation. *J. Mem. Lang.*, 31, 183–194.

Zwaan, R. A., and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychol. Bull.* 123, 162–185. doi: 10.1037/0033-2909.123.2.162

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 February 2016
doi: 10.3389/fpsyg.2016.00212

What's in a Name? Interlocutors Dynamically Update Expectations about Shared Names

Whitney M. Gegg-Harrison[1]* *and Michael K. Tanenhaus*[2]

[1] *Writing, Speaking, and Argument Program, University of Rochester, Rochester, NY, USA,* [2] *Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA*

In order to refer using a name, speakers must believe that their addressee knows about the link between the name and the intended referent. In cases where speakers and addressees learned a subset of names together, speakers are adept at using only the names their partner knows. But speakers do not always share such learning experience with their conversational partners. In these situations, what information guides speakers' choice of referring expression? A speaker who is uncertain about a names' common ground (CG) status often uses a name and description together. This N+D form allows speakers to demonstrate knowledge of a name, and could provide, even in the absence of miscommunication, useful evidence to the *addressee* regarding the *speaker*'s knowledge. In cases where knowledge of one name is associated with knowledge of *other* names, this could provide indirect evidence regarding knowledge of other names that could support generalizations used to update beliefs about CG. Using Bayesian approaches to language processing as a guiding framework, we predict that interlocutors can use their partner's choice of referring expression, in particular their use of an N+D form, to generate more accurate beliefs regarding their partner's knowledge of other names. In Experiment 1, we find that domain experts are able to use their partner's referring expression choices to generate more accurate estimates of CG. In Experiment 2, we find that interlocutors are able to infer from a partner's use of an N+D form which other names that partner is likely to know or not know. Our results suggest that interlocutors can use the information conveyed in their partner's choice of referring expression to make generalizations that contribute to more accurate beliefs about what is shared with their partner, and further, that models of CG for reference need to account not just for the status of *referents*, but the status of *means of referring* to those referents.

Keywords: common ground, reference, perspective-taking, belief-updating, conversation

OPEN ACCESS

Edited by:
Roger P. G. Van Gompel,
University of Dundee, UK

Reviewed by:
Zenzi M. Griffin,
University of Texas at Austin, USA
Hannah Rohde,
University of Edinburgh, UK
William S. Horton,
Northwestern University, USA

****Correspondence:***
Whitney M. Gegg-Harrison
whitney.gegg-harrison@rochester.edu

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

Received: *03 August 2015*
Accepted: *03 February 2016*
Published: *26 February 2016*

Citation:
Gegg-Harrison WM and Tanenhaus MK (2016) What's in a Name? Interlocutors Dynamically Update Expectations about Shared Names. Front. Psychol. 7:212. doi: 10.3389/fpsyg.2016.00212

1. INTRODUCTION

One of the most basic things we do with language is refer to things in the world. When we say something like, "Can you bring me the ball?," we are using the definite noun phrase *the ball* to refer to a particular ball in the world, and we're hoping that our addressee will be able to execute the requested action. In order to successfully refer, speakers must choose a referring expression that can be understood by their addressee(s), and this requires speakers to take into account what is in *common ground* (CG): the knowledge that is shared between conversational partners. This is particularly relevant when we consider the many choices for definite referring expressions—we can refer via a possessive, as in *my ball*, a definite description, as in *the ball with smudges on it*, a pronoun, as in *it*, or a proper name, as in *Wilson* - each of which assumes a different knowledge and attentional state on the part of the addressee (Grosz and Sidner, 1986; Gundel et al., 1993; Roberts, 2004), and can reflect the status of the referent in the preceding discourse (e.g., Ariel, 1990; Lambrecht, 1994). For proper names in particular, speakers need to know not just that the referent itself is in CG, but that their partner knows it *by that name*. What sources of information are available and used as the basis for our beliefs about what someone else knows, such that we could use this information in guiding our choices about what to say?

A major debate within the literature on CG focuses on the presumed computational complexity of generating and using representations of our partner's knowledge during language production or comprehension. An influential account from Clark and Marshall (1978, 1981) suggested that interlocutors could rely on elaborate "reference diaries" in memory, which enable them to find instances of "triple co-presence" between the speaker, addressee, and referent, and thus safely assume that a particular referent is mutually known. However, these rich representations strike many as psychologically implausible. Some alternate proposals hold that we are by nature egocentric, and instead of using information about their partner's knowledge or perspective, interlocutors use a heuristic: they assume that their own perspective or knowledge can serve as a proxy for what their interlocutor will know (e.g., Keysar et al., 2000; Wu and Keysar, 2007). Under these accounts, taking a partner's perspective into account requires effortful adjustment and monitoring processes, after something has gone awry. Several studies do suggest we are susceptible to making errors about what others know (e.g., Fussell and Krauss, 1992; Epley et al., 2004; Birch and Bloom, 2007), and that in particular, we fall victim to a "Curse of Knowledge" effect: we systematically assume that people know what we know.

But other researchers, such as Brown-Schmidt and Hanna (2011), argue that CG information is one of many partial constraints on language processing and production, and that information about ground status can, at least in cases where the cues to it are strong enough, influence language processing and production from the earliest moments. Brown-Schmidt and Hanna (2011) give an excellent overview of this debate, and point to a number of studies in which there is solid evidence for the use of CG as a constraint in both comprehension and production (e.g., Nadig and Sedivy, 2002; Clark and Krych, 2004; Brown-Schmidt et al., 2008; Brennan and Hanna, 2009; Brown-Schmidt, 2012). Studies from Heller et al. (2012) and Gorman et al. (2013) on referring expression choice, which we summarize in the following section, also lend support to this view of CG. The experiments we present in this paper build off this earlier work to ask whether interlocutors are capable of using the information conveyed by their partner's choice of referring expression to update their beliefs about what their partner knows.

1.1. Previous Work Using Names to Study CG-Use in Production

Proper names are arbitrary labels that can only be understood if the addressee knows the link between the label and the referent, and as such, they are ideal tools for exploring the use of CG in production. By teaching overlapping but non-identical sets of names to partners, it is possible to set up situations in which a name is either *privileged* (known only to one of the partners) or *shared* (known to both partners); in order to refer felicitously, speakers should only use those names which they know to be shared. Wu and Keysar (2007) used this paradigm to test their hypothesis that speakers use an "information overlap" heuristic to estimate common ground; they argued that instead of tracking the ground status of individual items, speakers instead rely on their estimate of the overall overlap between their own knowledge and their partner's. Indeed, in the *high overlap* cases (where speakers learned most names with their partner, and only a few were learned alone), they found that speakers used more privileged names than in the *low overlap* cases (where only a few of the names were shared).

However, in their replication of the Wu and Keysar (2007) study, Heller et al. (2012) found that in those cases where speakers used names for privileged items, these names were almost exclusively uttered along with a description, in what they (following Isaacs and Clark, 1987) call the "Name+Description" (N+D) form. Speakers included information that was necessary for their addressees to successfully identify the referent; evidence from additional studies suggested that these descriptions were not simply added as a repair or as a result of miscommunication, but were planned as part of the utterance from the beginning. This suggests that speakers are quite sensitive to the knowledge of their addressees, and skilled at tracking which names are shared, and which are privileged, when the basis for the shared knowledge is shared learning experience. The Name-Alone (N) form is reserved for those items which the speaker believes to be shared, and the N+D form reflects either a belief that that the item is privileged, or a lack of certainty about the item's ground status. But why use the name at all, when the addressee does not know it? Heller et al. (2012) suggested that the use of the N+D form may reflect a teaching strategy; if the speaker believed that they may need to refer to the item on subsequent turns, then it makes sense to use the name along with a description, rather than just a description, in order to "teach" the name to the partner. However, Gorman et al. (2013) explicitly disincentivized teaching by informing participants that they would only see each item once, and found that instead of decreasing their use of the N+D form for privileged items, speakers increased it; post-test debriefings further suggested that these speakers did not believe they were teaching names to their addressee. Though this indicates that speakers were not strategically *teaching* the names to their addressee via the N+D form, this does not mean that addressees would have been incapable of *learning* something from the speakers' use of the N+D form; we will return to this possibility shortly.

Building off the work of Heller et al. (2012), we conducted a series of studies (described in Gorman et al., 2013) exploring the memory representations that support CG use during language production. We based our approach on the framework presented by Horton and Gerrig (2005a,b), who propose that information about CG is represented as a by-product of ordinary memory processes, which contain context-specific episodic traces. Results from the spoken word recognition literature suggest that people might indeed have automatic access to speaker-specific episodic traces for names (Goldinger, 1998; Creel et al., 2008; Creel and Tumlin, 2011). Work on lexical precedents from Metzing and Brennan (2003) and Brown-Schmidt (2009) also demonstrates that addressees can use speaker-specific information when comprehending referring expressions (but cf Kronmüller and Barr, 2015). In short, it seems the representations upon which language use depends (e.g., word representations) already encode speaker-specific information. This might explain how such purportedly rich CG representations as the ones described by Clark and Marshall (1978, 1981) could be used during real time conversation. Our conversational partner and everything about the context in which we are speaking to them serve as cues that make associated information more accessible in memory. One major question, then, is precisely what kind of information needs to be accessible in memory. When we decide we want to refer to something, it seems we need information about whether that referent is "shared" with our addressee (to support what Horton and Gerrig, 2005a would call "commonality assessment"), but also about whether a particular *means of referring* to that referent is likewise "shared" (to support what they would call "message formation"). That is, rather than "triple co-presence," it seems we need evidence for a sort of "quadruple co-presence": evidence that we, our addressee, the referent, and the means of referring to it have all been "co-present." Our work has been aimed at probing the factors that can support inferences regarding the status of particular means of referring to particular referents with a particular conversational partner.

In the studies reported in Gorman et al. (2013), participants learned novel names for novel creatures during the training phase, then interacted during the testing phase in a referential communication game using those creatures. One participant was named the Director, and either learned shared names alongside their partner (the *together* condition), or learned them separately but were told that their partner learned those same names (the *alone* condition). In both conditions, the Director went on to learn a set of privileged names that were not learned by their partner. These studies found support for the Horton and Gerrig (2005a,b) claim that shared experience should enable the development of episodic memory traces linking the speaker, the addressee, and the referents, and thus support the use of CG during production: Directors were far better at avoiding the use of the N form for privileged items in the *together* condition than in the *alone* condition, where such episodic memory traces would not be present. However, even in the *alone* condition, Directors were still much more likely to use the N form for shared items than for privileged items; even without the benefit of shared experience, Directors were still able to use what they had been told about their partner's knowledge, though not nearly as successfully as when the shared knowledge was established through shared experience.

Interestingly, Directors in the *alone* condition were also more likely to use the N+D form for privileged names than Directors in the *together* condition, suggesting that use of the N+D form may reflect greater uncertainty about the ground status of items. One possibility is that the collaborative learning in the *together* condition aided speakers because it provided better context cues to distinguish between shared and privileged information, not just because of partner-specific episodic memory cues. For the Directors in the *alone* condition, almost nothing distinguishes shared and privileged names in memory, since both are learned in isolation. In contrast, in the *together* condition, many pairs collaboratively created memory cues to help remember the names, and the context in which shared names were learned (interacting with another person and the experimenter) and the context in which privileged names were learned (sitting alone with the experimenter) were quite different. As such, another study reported in Gorman et al. (2013) aimed to explore whether the relevant memory cues depend on *partner-specific* shared learning. A *third-party* condition was introduced, in which the Director learned shared names together with a partner, but this partner was not the same partner with whom they would interact in the referential communication task; the Director was simply told that their new partner had learned the same names as their earlier partner. Thus, the shared names were still established via shared experience, but that experience then needed to be generalized to the new partner. It was found that in the *third-party* condition, Directors did nearly as well at avoiding the use of the N form for privileged items as Directors in the *together* condition, but were far more likely to use the N+D form for both shared and privileged items than Directors from either of the other conditions. Directors in the *third-party* condition were able to generalize their shared experience with a third party to their new partner, but were again left with greater uncertainty about the ground status of individual items.

While partner-specific episodic memory traces may be a powerful influence on speakers' ability to use information about the ground status of an item and its name, they are clearly not the only information available to speakers; simply being told what another person has learned was enough to generate at least some use of ground information by the speakers, and learning shared vs. privileged names in more distinguishable contexts helps. This raises a question as to what kinds of representations speakers might be creating in order to accurately remember (or generate expectations about) what names are shared with a particular addressee and to use them (relatively) appropriately, and what kinds of information allow speakers to successfully generalize beyond their direct shared experience with a conversational partner. This is a particularly important question to investigate: we often interact with people with whom we do not share experience, but do share knowledge. Thus, we need to be able to go beyond the direct evidence that comes from shared experience—to make inferences about our partner's knowledge, and generalize beyond those inferences. Our results from Gorman et al. (2013) suggest a potential basis for inferences about shared knowledge based on common community membership (Clark and Marshall, 1978, 1981). Speakers are more likely to use the N+D form when they do not share learning experience with their partner; this allows them to use a name without assuming that their partner knows the name. The speakers' use of this name, in turn, could serve as a cue to the addressee that the speaker is indeed a member of the same community, and thus support a generalization: that the speaker knows other names associated with that community. This chain of inferences and generalizations could support more accurate estimates of CG information over the course of conversation—even when there is no miscommunication between interlocutors.

1.2. Motivation for Current Experiments

Recall that the strongest evidence used to support the argument that speakers are egocentric comes from demonstrations of the "Curse of Knowledge": the seemingly irrational belief that others know what we know. But is this really so irrational? Consider the task facing speakers interacting with a partner with whom they don't share learning experience: they must adapt to their conversational partner's knowledge in order to refer successfully. Speakers typically interact with partners who are relatively similar to themselves (this is especially true in experiments that involve pairs of college students). In Bayesian terms, people's apparent initial egocentricity may be the result of strong prior expectations that their partner knows what they know (perhaps tempered by the degree to which they see their partner as similar to themselves, and their prior experiences with similar conversational partners and similar conversational contexts). By starting with their own knowledge as a prior estimate of their partner's knowledge, speakers are arguably behaving more rationally than if they used a completely unbiased prior estimate.

Over the course of conversation, interlocutors will be exposed to evidence regarding their partner's knowledge that they could use to update their expectations about what is shared. Recall that Heller et al. (2012) and Gorman et al. (2013) showed that a speaker who is uncertain about a names' CG status often uses the N+D form. This form allows the speaker to demonstrate knowledge of a name without making assumptions regarding whether that name is shared. It thus could provide useful evidence to the *addressee* regarding the speaker's knowledge—and this evidence is available even in the absence of miscommunication. This type of evidence could be particularly useful as a cue to common discourse community membership, if knowledge of one name is associated with knowledge of other names (as in domains where varying levels expertise are associated with use of particular names). Do interlocutors attend to and use this evidence to rationally update their expectations about their partner's knowledge? We explored this hypothesis in two experiments where choice of referring expression could serve as a cue to knowledge of other names. In Experiment 1, we find that domain experts are able to use their partner's referring expression choices to generate more accurate estimates of CG. In Experiment 2, we find that interlocutors are able to infer from a partner's use of an N+D form which other names that partner is likely to know or not know.

2. EXPERIMENT 1: CG BELIEF-UPDATING IN UNSCRIPTED TASK-ORIENTED DIALOG

In Experiment 1, we embedded the task of name learning in the context of a rich "toy" world by creating a role-playing game in which certain levels are always encountered before others, and the participant's choices can make regions of the world (and the information contained there) inaccessible; this makes it possible that a speaker displaying knowledge of one name can implicate that they also have knowledge of the names learned prior to that one in the game. We hypothesized that domain experts could use their partner's referring expression choices to update their beliefs regarding their partner's domain-specific knowledge. We asked the following questions: would Expert's initial beliefs about shared knowledge reflect partner-specific information when available? Could the partner's referring expression choices provide a useful cue to the set of names known by that partner? Would Experts adjust their beliefs about partner knowledge on the basis of their partner's referring expression choices? And could Experts generalize, inferring that one name is known given a display of knowledge of another name?

2.1. Methods

2.1.1. Participants

Two native English speakers from the University of Rochester were recruited to serve as Game Experts and were paid hourly rates for their participation. A further 32 native English speakers from the University of Rochester were paid to participate as novice gamers. Four naive participants were brought in for each 2-day experimental session. All participants signed a written consent form which was approved by the Research Subjects Review Board of our institution.

2.1.2. Materials

Three novel clipart images of "cute monsters" from clipart.com were modified to create nine unique creatures: 3 Wugs, 3 Lorks, and 3 Greps. The individual Wugs, Lorks, and Greps were distinguished by a "feature" designed to resemble a rune, created using the paintbrush in GIMP, drawn on the creature's belly; each feature was assigned an invented name that was one (CVC) syllable in length, and designed to be easily distinguishable from other feature names. Each creature was assigned an invented name (e.g., "Gramperoo"). These creatures were presented over the course of the game. **Figure 1** illustrates three sample creatures and all six features used in the game.

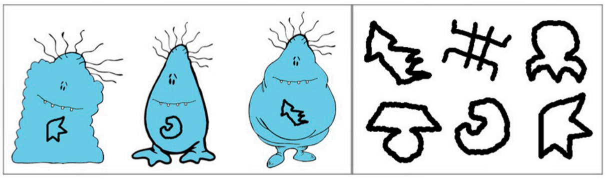


FIGURE 1 | Experiment 1: example creatures and features.

2.1.3. Procedure

2.1.3.1. Expert training

In order to create Experts who had full knowledge of everything in the game, each Expert was given a copy of the Game Book (which contained the "story" of the game, along with names and other information about each creature within the game) and specific training. On Day 1, the experimenter took the role of guide, and led the Experts through the game as naive participants. Then the Experts were given 3 weeks to study the Game Book. Experts were quizzed weekly on their memory for the names and other information contained in each level of the game. By the end of training, Experts completed these quizzes with no errors.

2.1.3.2. Day 1

In order to create a situation through which Experts could develop structured shared knowledge with participants, each Expert guided two participants through the role playing game, one at a time. The Expert read the story from the Game Book to the participant, told the participant when to move their game piece on the Game Board, informed the participant of the result of dice rolls and gift choices, and quizzed them on names and other information at the relevant points during the game. Participants played the game twice with the Expert; most did not make it to the final stages of the game.

2.1.3.3. Day 2

A new set of Expert-participant pairings was created: each Expert interacted with one of the participants they'd guided the previous day, and one of the participants the other Expert had guided. **Table 1** illustrates the pairings; the order was randomized (Participant A was not always first). This allowed us to explore both partner-specific knowledge and inferences rooted in general expectations for what an unfamiliar game-player might know.

The Experts and Participants completed the following series of tasks on Day 2: a Pre-Test, a Matching Task, a Mid-Test, a SET game, and a Post-Test.

2.1.4. Pre-Test

2.1.4.1. Expert

In order to probe the Expert's expectations about their Day 2 partner's knowledge, the Expert was given a worksheet with images of each of the runes and each creature and marked a spot on an 11cm line representing their belief (from "most likely no" to "most likely yes") regarding the likelihood that the Day 2 partner knew the name. The location of the marks were later measured using a ruler and recorded in a spreadsheet.

TABLE 1 | Expert-participant pairings.

Day 1		Day 2	
Expert 1:	**Expert 2:**	**Expert 1:**	**Expert 2:**
Participants A and B	Participants C and D	Participants A and C	Participants B and D

2.1.4.2. Participant

In order to establish what the naive participants actually remembered from their Day 1 experience, the naive participants were given a separate worksheet containing images of each of the runes and each creature. They indicated whether they had learned each name, and wrote down the name as they remembered it.

2.1.4.3. Matching Task

The Expert and participant completed a Matching task using cards printed with images of each of the creatures learned in the game, as well as three novel creatures, created using the three family body-shapes, but with rune-markings on the belly that did not correspond to any of the learned characters. The participant and Expert sat in the same room with their backs to one another, so that they could hear each other speaking but could not see the each other's cards. The experimenter placed the cards in a specific order on the participants' table, and then placed the corresponding cards on the Expert's table, in a different specific order. The participant was told to work with the Expert until the cards on the Expert's table were in the same order as the cards on their own table, and the pair was encouraged to converse freely as they worked to accomplish the task. The Matching Task was chosen to provide an opportunity for the naive participant to use the names (and thus provide evidence about their knowledge to the Expert). Their conversation was recorded and transcribed. Transcriptions were later annotated by the experimenter, who marked each reference to a character or rune and tagged it as belonging to one of three categories: Name Alone (N), Name + Description (N+D), or Description Alone (D).

2.1.5. Mid-Test

2.1.5.1. Expert

In order to assess changes in the Expert's beliefs regarding their partner's knowledge following the Matching Task, the Expert was given a second copy of the worksheet they completed in the Pre-Test to complete after the Matching task.

2.1.5.2. Participant

The naive participant was given a new worksheet, on which they answered questions regarding the difficulty of the Matching task, and their strategy for referring to creatures they knew and creatures they did not know.

2.1.5.3. SET task

The participant and Expert were seated as in the Matching Task. Each set of cards from the Matching Task were shuffled so that they were in random order; each set was arranged in 3 rows of 4. The Expert and participant were told to work together to form "sets" of cards that shared some common characteristic (e.g., physical appearance, the jobs or territories of the creatures, or the sounds of the creatures' names). The SET task was chosen as a "targeted language game" (Brown-Schmidt and Tanenhaus, 2008; Tanenhaus and Brown-Schmidt, 2008), designed to elicit conversation regarding what each participant knew about the creatures. The Expert and participant took turns choosing two cards from their set of cards; their partner's job was to choose a card from their set of cards that would complete a set with the first two cards that were chosen. They were encouraged to choose their two cards carefully when it was their turn, so that their partner had the best chance of being able to complete a set. The Expert and participant were allowed to tell their partner what characteristic they had in mind for the set, but could not coach their partner on which specific card to use. The Expert and participant were each given one "PASS" to use if they could not complete a set, and were told the goal was to use as many cards as they could before using their PASSes. The conversations and card choices were recorded and transcribed. The transcriptions were later annotated by the experimenter, who marked each reference to a character or rune and tagged it as belonging to one of 3 categories: N, N+D, or D.

2.1.6. Post-Test

2.1.6.1. Expert

In order to assess changes in the Expert's beliefs following the SET task, the final worksheet asked Experts to give a final YES/NO judgment regarding their partner's knowledge of names, and also asked about changes in their beliefs about their partner's knowledge from the game over the course of completing the two tasks, and their strategy during the SET task.

2.1.6.2. Participant

The final worksheet asked questions regarding the difficulty of the SET task and their strategies during the SET task, as well as questions about their memory for the names of the creatures and runes, how often they thought they'd used the names, and their strategy for referring to creatures or runes whose names they did not know.

2.2. Experiment 1 Results

2.2.1. Experts' Initial Beliefs

We first assessed the Expert's initial beliefs and explored the extent to which these beliefs relate to the knowledge of their Day 2 partner. We converted the Experts' Pre-Test number-line ratings into YES/NO judgments by norming them to a value between 0 and 1, and assigning estimates below 0.5 to NO and estimates above 0.5 to YES. If the partner had answered "no" or gave an incorrect name for an item in the Pre-Test, they were coded as not knowing the name; if they answered "yes" and gave a correct name for an item, they were coded as knowing the name. Experts were correct in their Pre-Test judgments about which names their partner knew and did not know 80% of the time when they were working with the same partner as on Day 1, and 68% of the time when they were working with a different partner than on Day 1, which suggests some use of partner-specific information. **Table 2** compares the expectation of the Expert regarding whether their Day 2 partner knew a particular name with the actual knowledge of that partner, as indicated by the partner's Pre-Test responses. Incorrect judgments are bolded. Note that there does appear to be a "Curse of Knowledge" effect in the Experts' response patterns, particularly when the Expert is working with a different partner on Day 2 than on Day 1: Experts assume their partner knows a name when it is not actually known 27.6% of the time when working with the partner, and 42.7% of the time when working with a different Day 2 partner.

However, Experts' basis for their beliefs regarding partner knowledge is more likely to be the experience in the game itself on Day 1, during which their partners may have learned names that were subsequently forgotten by Day 2, and thus in **Table 3**, we present the percentages of correct and incorrect judgments about partner knowledge broken down by what their partner actually learned, rather than what they report remembering on Day 2; note the dramatic reduction in incorrect judgments about knowledge status for items that the Expert expects to be known (from 27.6 to 2% for the same partners, and from 42.7 to 11.7% for different partners).

To test whether Expert's judgements reflect parter-specific information, we modeled the accuracy (based on their partner's Day 2 Pre-Test knowledge) of Experts' Pre-Test Yes/No judgments about whether a particular item was known using a mixed effects logistic regression model with the partner's status (same as Day 1 or different) as a fixed effect, along with random effects for the partner, Expert, and item. We found a significant main effect of the partner's status, such that Experts were more likely to be accurate in their judgments about whether their partner knew a particular item if their partner was the person they had played the game with on the previous day ($\beta = 0.82$, S.E. $= 0.39$, $p < 0.05$). But when we add Game Experience (whether or not a name was learned by the partner on Day 1) to this model, the partner's Day 2 knowledge is no longer a significant predictor; instead, we find a main effect of Game Experience, such that Experts' ratings of the likelihood that their partner knows a name are significantly higher when that name was learned during Day 1 ($\beta = 1.16$, S.E. $= 0.38$, $p < 0.01$), and there is no interaction with partner status, likely due to the

TABLE 2 | Expert's Pre-Test judgments of what Day 2 Partner knows and does not know, compared to the Partner's actual knowledge (based on Pre-Test).

	Knows (%)	Doesn't Know (%)
SAME PARTNER		
Expert expects Known (33.8% of responses)	72.4	**27.6**
Expert expects Unknown (66.2% of responses)	**15.2**	84.8
DIFFERENT PARTNER		
Expert expects Known (15.5% of responses)	57.3	**42.7**
Expert expects Unknown (84.5% of responses)	**28.6**	71.4

TABLE 3 | Expert's Pre-Test judgments of what Day 2 Partner knows and does not know, compared to the names Partner actually learned on Day 1 (Game Experience).

	Learned (%)	Didn't learn (%)
SAME PARTNER		
Expert expects Known (33.8% of responses)	98	**2**
Expert expects Unknown (66.2% of responses)	**49.4**	50.6
DIFFERENT PARTNER		
Expert expects Known (15.5% of responses)	88.3	**11.7**
Expert expects Unknown (84.5% of responses)	**48.2**	51

fact that most participants had relatively similar performance (and thus learned similar names) during Day 1. Thus, while Experts' judgments regarding the likelihood that their partner knows a name reflect partner-specific information, they are still relatively accurate for new partners, based on their implicit sense of which names were likely to have been learned during the game.

2.2.2. Patterns of Referring Expression Choice during Matching Task

In order to explore whether the partner's use of referring expressions could have provided a useful cue to the Expert regarding that partner's knowledge, we annotated the transcripts from the Matching task, coding each reference to a named feature or creature as either an N, an N+D , or a D. If a creature was referred to using a description that included the feature name (e.g., "The Wug with the Bor" in reference to the creature *Molgiroo*, a "Wug" family creature with the "Bor" feature), this was coded as an N for the feature and a D for the creature; features could also be described rather than named, as in "The Wug with the thing that looks like a rocket," and this would be coded as a D for both the feature and the creature. Because of the nature of the matching task, participants were sometimes able to complete the task without referring to all of the creatures; these were coded as "NONE." We then recorded the final utterance type from the naive participant referring to that feature or creature (N, D or N+D). This gave us a measure of the evidence provided by the participants' utterances. **Figure 2** shows the distribution of the partners' first utterances during the Matching task for known and unknown creatures and features.

For items whose names the partner did know, according to the partner's Pre-Test, the partner used the N form 37% of the time, the N+D form 9% of the time, and the D form 48% of the time.

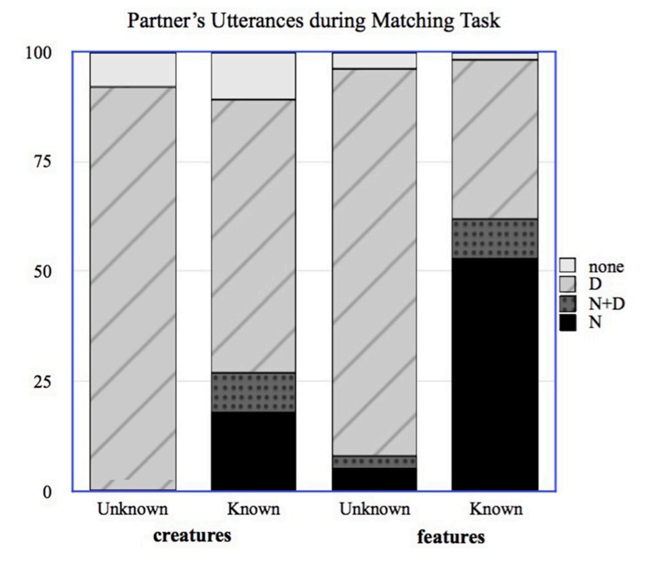


FIGURE 2 | Distribution of naive participants' utterances for known and unknown creatures and features in the Matching task.

The vast majority of names used by the partner were for features; 81% of the partner's N uses and 53% of the partner's N+D uses were for features rather than for creatures. Note that the partners here are using the N form more often than the N+D form for the names they know; this is likely due to the fact that they are interacting with people whom they know to be Experts, and so can safely refer using only the name. Experts almost never used names during the Matching Task, which was almost certainly driven by the fact that they were playing the role of matcher rather than director.

Note that there are a small percentage of N and N+D uses by the partner for "unknown" items. In all cases where the participant used the N form for an "unknown" item, the participant initially used an incorrect form of the name (and had done so during the Pre-Test as well), and the Expert provided a correction, which the partner then proceeded to use throughout the task, as in the following example:

Partner: The next one has a **Rep** on it...
Expert: *Do you mean* **Rab***? The half-star?*
Partner: *Oh yes, sorry,* **Rab***!*

In sum, partners did not use names for all of the items whose names were known to them during the Matching task, but they did use some names. In choosing to use names for particular items, partners may have provided evidence to the Expert regarding their knowledge not only of that item, but of other items that should have been learned alongside it. Likewise, in deciding to use a description for an item, partners may have provided evidence to the Expert that suggested a lack of knowledge of that item's name, evidence that could be misleading if the name is actually known by the partner. Did this evidence contribute to changes in Experts' beliefs about partner knowledge? In the next section, we provide an overview of the changes in Experts' beliefs about what their partner did and did not know, and evaluate the extent to which these changes can be predicted by the evidence provided by the partner during the Matching Task.

2.2.3. Changes in Experts' Beliefs at Mid-Test

In order to explore the changes in Experts' beliefs about partner knowledge and the extent to which these changes were driven by the evidence provided in the Matching Task, we first converted the Experts' likelihood estimates from the Mid-Test into YES/NO judgments in the same manner as for the Pre-Test estimates. We find that in the Mid-Test, the difference between Experts' accuracy on YES/NO judgments for same Day 2 partner vs. different Day 2 partner disappear: Experts were correct in their judgments around 79% of the time for both types of partner, meaning that Experts' accuracy for different Day 2 partners improved by 10% over their performance in the Pre-Test.

A summary of the Experts' judgment data from the Mid-Test, broken down by the items the partner did and did not actually know, is given in **Table 4**.

An examination of **Table 4** in comparison to **Table 2** reveals some interesting changes. Most striking are the improvements

TABLE 4 | Experts' Mid-Test Judgements relative to Partners' Pre-Test Memory for names.

	Knows (%)	Doesn't Know (%)
SAME PARTNER		
Expert expects Known (22.3% of responses)	80.7	**19.3**
Expert expects Unknown (77.7% of responses)	**21.1**	78.9
DIFFERENT PARTNER		
Expert expects Known (14.6% of responses)	90.6	**9.4**
Expert expects Unknown (85.4% of responses)	**23.2**	76.8

in Experts' judgments regarding the knowledge of their partner when their partner was different on Day 2; for different partners, Experts correctly believe things to be known when they are in fact known 90.6% of the time in the Mid-Test, compared to 57.3% of the time in the Pre-Test. The "Curse of Knowledge"-type errors are reduced when working with the same partner, as well: Experts correctly believe things to be known when they are in fact known 80.7% of the time in the Mid-Test compared to 72.4% of the time in the Pre-Test.

In order to assess whether these differences reflected significant improvements in judgements from the Pre-Test to the Mid-Test, we used a mixed effects logistic regression model to predict whether the Yes/No value of the Expert's judgment was correct, with the test from which that judgment came (*Pre-Test* or *Mid-Test*) and the partner status (*same Day 2* or *different Day 2*) as fixed effects and participant number, item name, and Expert name as random slopes and intercepts. We found a significant main effect of judgment type on accuracy, such that judgments that came from the Pre-Test were less likely to be accurate than judgments that came from the Mid-Test ($\beta = -0.58$, S.E. $= 0.18$, $p < 0.01$).

To test whether Experts' Mid-Test judgments were influenced by the evidence provided in the form of their partner's referring expression choices, we used a mixed effects multi-level regression model to predict the Expert's Mid-Test Yes/No judgment for each item, with the referring expression choice of the partner for that item (N, N+D, D, or none) as a fixed effect and participant number, item name, and Expert name as random slopes and intercepts. Experts were significantly more likely to rate an item as "known" when the partner had used the N form to refer to the item ($\beta = 6.45$, S.E. $= 0.82$, $p < 0.0001$) and also when the partner had used an N+D form ($\beta = 3.97$, S.E. $= 0.86$, $p < 0.0001$), and were significantly less likely to rate an item as "known" when the partner had used a description ($\beta = -4.04$, S.E. $= 0.52$, $p < 0.0001$).

We had hoped to test whether Experts would alter their belief about the status of one item based on the evidence provided in relation to another item. However, the patterns of referring expression choice by the partners made that impossible; there were not enough cases where a partner used a name from "later" in the game without also having used one from earlier in the game. There was one striking example, in which the Expert was working with a different partner on Day 2 than he had worked with on Day 1. This partner had made it to the end of the game on Day 1, and during the Matching Task, used names for some, but not all of the creatures she had encountered. She did, however, refer to the final creature from the Day 1 game as "King Floogelor" during the Matching Task, and the Expert reacted with surprise that she knew that name. In this particular case, the Expert dramatically increased his ratings on the Mid-Test (compared to the Pre-Test) for all items. But this was the only example of this type in the dataset. In order to ask specifically about whether the use of one name can lead to generalizations about knowledge of other names, it may be necessary to use partially-scripted games; we present one such approach in Experiment 2.

2.2.4. Patterns of Referring Expression and Set Choice during SET Task

In order to explore whether these changes in beliefs about partner knowledge were reflected in the Expert's referring expression choices during the SET task, we used the same annotation and coding scheme as for the transcripts for the Matching Task, additionally coding the first utterance type from the Expert. **Figure 3** gives the distribution of utterance types (N, ND, D, and none) during the SET task for partners and for Experts, respectively.

Unlike the Matching task, Experts used many names during the SET task, restricting their use of names to those items that were actually known by their partner; this was aided by their updated beliefs regarding their partner's knowledge. Models using the Expert's Mid-Test judgments to predict the choice of the N-form in the SET task are a better fit (based on AIC) than those using the Expert's Pre-Test judgments. The connection to the evidence provided by the partner is strong: when we use the form of referring expression chosen by the partner in the Matching Task to predict the form of referring expression chosen by the Expert in the SET Task, we find that Experts are significantly more likely to use the N-form when the partner used the N-form ($\beta = 4.94$, S.E. $= 0.72$, $p < 0.001$) or the N+D form ($\beta = 3.27$, S.E. $= 0.95$, $p < 0.001$) for that item. The handful of instances in which an N form was used for an unknown-to-the-partner item occurred on the final turns of the SET game; at that point it was obvious which card remained, and so the utterance could be understood by the partner even without knowing the name.

2.3. Experiment 1 General Discussion

Experts' initial beliefs regarding their Day 2 partner's knowledge of names did reflect partner-specific information, as evidenced by the higher accuracy of their Pre-Test judgements when working with the same Day 2 partner as when working with a different Day 2 partner. When working with unfamiliar partners on Day 2, Experts' initial beliefs regarding those partner's knowledge was influenced by the Experts' game-playing experience, and thus still more accurate than might be expected if the Experts' had no basis for forming expectations regarding their partners' knowledge. The partner's referring expression choices during the initial referential communication task did provide useful information regarding the partner's knowledge of names, but partner did not *always* use the N form for names that were known. Experts were

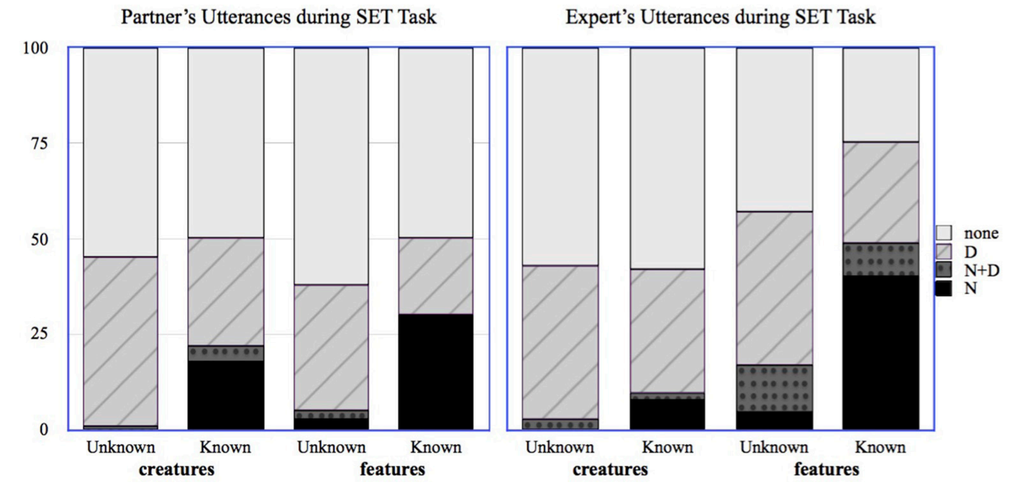


FIGURE 3 | Distribution of partner's and Expert's utterances for known and unknown creatures and features in the SET task.

able to use the information provided by the form of their partner's referring expressions to generate more accurate beliefs regarding their partner's knowledge of names, which was apparent both in Mid-Test and Post-Test explicit judgements, and in the form of referring expression chosen by the Expert in the final SET task.

There are limitations inherent to the design of this study. Because we used only two Experts, it is difficult to draw general conclusions about what people with expertise do when speaking to those whose knowledge only partly overlaps with their own. Another concern is that by explicitly asking the Experts to make repeated judgments about their partner's likely knowledge, we highlighted the issue of partner knowledge for the Experts in a way that typical conversation does not, and thus made that information more available and salient, which Galati and Brennan (2010) argue is a key factor for finding evidence of CG-use; over the course of the experiment, each Expert was asked to complete 16 Pre-Tests, 16 Mid-Tests, and 16 Post-Tests regarding their partners' knowledge. We can't argue that these tests did not bias the Experts toward more careful attention to their partner's knowledge state, but the results of Experiment 1 do show that it is *possible* for people to attend to the evidence provided by their fellow interlocutor in the form of their choice of referring expression in order to update their beliefs regarding what is and is not shared, and can use this information in deciding how to refer in subsequent conversation.

What Experiment 1 could not reveal is whether interlocutors were capable of generalizing on the basis of their partner's displayed knowledge. In Experiment 2 we explored this possibility using a partially-scripted online game 2 that specifically promotes the use of the N+D form.

3. EXPERIMENT 2: CG BELIEF-UPDATING IN A PARTIALLY-SCRIPTED DIALOG TASK

In Experiment 2, we investigated whether a partner's use of a name could lead an interlocutor to generalize about the other knowledge that partner may have. We developed a simplified name-learning game that shared some critical properties with the game in Experiment 1. This game was posted as a "HIT" on Amazon Mechanical Turk, allowing us to obtain data from a larger population. Participants learn the names of creatures while solving increasingly more challenging timed math problems, and then must choose between the Red Path or the Blue Path, and once they choose, they cannot learn names on the alternate path. Thus, participants' knowledge at the end of the game varies, depending both on their ability to solve math problems, and on their choice of path if they make it far enough into the game—a situation that mirrors the one created by the dice rolls and choice points in Experiment 1. Following the name-learning game, participants were asked to take part in a referential communication game with another participant from the name-learning game. This second game player was actually an automated agent, who we will call *AutoTurk*, programmed with particular experiences from the game: in one condition (*RedExpert*), AutoTurk was an expert who made it all the way to the end of the Red Path; in another (*BlueExpert*), AutoTurk made it all the way to the end of the Blue Path; and in the third (*EarlyFailure*), AutoTurk was a poor player, who failed out of the game by solving a math problem incorrectly after learning only two character names. We collected data regarding how participants shifted their expectations regarding partner knowledge during the course of conversation, both via explicit judgments and via their referring expression choices. We hypothesized that upon hearing a partner use a name for an item from the Red Path, participants would be more likely to believe their partner knows other Red Path names, and *less* likely to believe their partner knows Blue Path names.

3.1. Methods

3.1.1. Participants

Hundred and twenty naive adult speakers of English volunteered to participate in the study for payment via Amazon Mechanical

Turk. Prior to accepting the Mechanical Turk "HIT," participants gave consent via a digital consent form approved by the Research Subjects Review Board of our institution.

3.2. Materials

Three novel clipart images of "cute monsters" from clipart.com were modified to create seven unique creatures from three families, each with invented names, as in Experiment 1. These creatures were presented over the course of game, whose layout is depicted in **Figure 4**. A template was used to generate five unique math problems for each participant. They were asked to solve the problems to progress through the game; each "splat" symbol on the game paths represents a math problem.

3.3. Procedure

3.3.0.1. Instructions

The participant, prior to accepting the HIT, read instructions describing the game, its rules and payout structure. Each participant was informed that they would be playing a game that involved two separate stages: learning the names of the cartoon creatures while solving math problems, then playing a matching game with a networked partner who had also completed the first stage. They would acquire points for each correctly learned name and correctly solved math problem during the first stage; these points would carry over into the second stage, where they would acquire points based on how well they solved the matching game with their partner. Participants were also told that if they successfully completed Stage 1 (by making it to the end of either path), they would receive a bonus payout ("Bonus 1"); another bonus payout ("Bonus 2") was based on the total number points across Stage 1 and Stage 2 combined.

3.3.0.2. Stage 1

The participant was presented with the game screen, and introduced to the first character, as depicted in **Figure 4**, and told to remember its name. They were then given 8 s to solve a simple addition problem; if successful, they advanced to the next round of the game, and were introduced to another creature. The math problems were designed to have the form X + Y, where X and Y were both single digit numbers for the problems that were solved prior to the path-split. For the last two problems, solved after the path-split, X and Y were both double digit numbers if the participant had chosen the Red Path, while only one of X or Y was a double-digit number if the participant had chosen the Blue Path. Participants were made aware that the choice of the Red Path would entail more difficult math problems, but a higher Bonus 1 payout. If the participant successfully reached the "midpoint" of the game (after learning the third creature), they were given a Mid-test: they had to choose each creatures' name from a list of four possibilities. If successful, they were given a choice between the Red Path and the Blue Path. If the participant successfully learned both names on their chosen path, they were given a final test in which they had to again choose each creatures' name from a list of four possibilities before they could complete the path and advance to the Practice Phase. If the participant incorrectly solved a math problem or failed to successfully complete the Mid-test, their time in Stage 1 ended, and they were advanced into the Practice Phase—this allowed us to create a believable scenario in which a random participant in the game could know anywhere from 0 to 5 names.

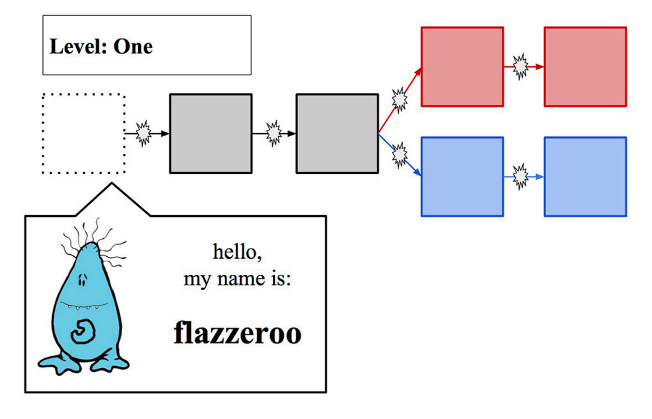


FIGURE 4 | Game Layout/Introduction of First Creature; each box represents a character whose name could be learned, while each "splat" represents a timed math problem to be solved. The top path ("Red Path") had more challenging math problems and a higher payout.

3.3.0.3. Practice phase

To ensure that participants remembered the names they had just learned, they were next presented with a series of single creatures in random order, as well as a list of 10 possible names (three were "distractor" names that did not belong to any creature), along with the options "did not learn" and "do not remember." For each creature, the participant needed to correctly identify its name if it was a creature they had learned, or correctly identify it as an unlearned name if it was a name they had not learned. If the participant chose an incorrect name, or chose "do not remember" for a name they had learned, they were reminded of the creature's name; if they chose any name at all for a creature they had not learned, they were reminded of the fact that they did not learn it. For each creature, the participant needed to correctly identify its name (or status as an unlearned creature) twice before advancing to Stage 2.

3.3.0.4. Stage 2

Participants were told that they were being connected to another Mechanical Turk "worker" who had also participated in Stage 1 of the game, and were shown a "wait" symbol and a progress bar, which changed to a "connection" symbol after a random time lag of between 2 and 90 s; in reality, they were "connected" to *AutoTurk* (the automated agent described earlier). Then, participants were asked to give YES/NO judgments regarding their expectation that their random partner would know each of the seven creatures. Next, participants were told they would be playing a matching game with their partner (who was actually AutoTurk). When it was the participants' turn to be the director, they were shown a single creature; their job was to decide how to refer to that creature so that their partner (actually AutoTurk) could correctly identify it from an array of four creatures.

On director trials, participants were presented with a list of 10 names and a list of 10 descriptions from which they could choose using radio buttons. They were told that if their partner chose the correct creature based on what they said, they would receive 8 points by default; if they used a name along with a description, they'd receive 5 bonus points, and if they used just the name, they'd receive 10 bonus points. If their partner chose the wrong creature based on what they'd said, they'd lose 10 points (Thus, using the name by itself if the partner does not know it would result in losing 10 points instead of gaining 18). The participant took turns with AutoTurk playing the role of director or matcher; when the participant was the matcher, they saw four creatures on the screen, and were presented with the referring expression their partner (actually AutoTurk) had selected: either N, N+D, or D. When the participant was the matcher, they received 10 points for selecting the correct item based on AutoTurk's utterance, and lost 10 points if they chose the wrong one.

Critically, the trials (shown in **Table 5**) were ordered such that it was possible for the AutoTurk to use either an N+D or a D form for particular creatures, which could then serve as a cue to the participant about whether their partner knew those creatures; in Trial 4 (a Blue Path creature), *BlueExpert* uses an N+D while *RedExpert* uses a D, and in Trial 6 (a Red Path creature), *BlueExpert* uses a D while *RedExpert* uses an N+D. Since participants should only choose the N form for a particular creature if they believed their partner actually knew the name, this allowed us to use subsequent choices between N and N+D for creatures the participant knew as a measure of their beliefs regarding their partner's knowledge.

3.3.0.5. Post-Tests

To allow us to track whether participant's explicit judgements of their partner's knowledge shifted as a result of playing the referential communication game, the participant again made Yes/No judgments regarding their belief that the partner they had worked with knew each of the creatures. They were also asked whether they had paid attention to whether their partner used names, and were asked to describe their own strategy for completing the game in Stage 2, and what strategy they thought their partner was using.

TABLE 5 | Experiment 2: Stage 2 Trials.

Trial	Utterance Form/AutoTurk Knowledge		
	RedExpert	BlueExpert	EarlyFailure
1. Flazzeroo		Participant's Choice	
2. Floogirep	N+D	N+D	D
3. Gramperoo		Participant's Choice	
4. Bampirep (Final Blue)	D	N+D	D
5. Molgirep (Blue)		Participant's Choice	
6. Narpelor (Final Red)	N+D	D	D
7. Trimmelor (Red)		Participant's Choice	
8. Flazzeroo	N	N	N
9. Narpelor (Final Red)		Participant's Choice	
10. Gramperoo	N	N	N

3.4. Results

3.4.1. Explicit Judgments of Partner Knowledge

One issue worth noting is that 7% of participants, in the Post-Test, gave Yes judgments to all of the creatures; they indicated that they believed the partner they had worked with knew all of the creatures from the Red Path and from the Blue Path, which is impossible. This suggests that at least some participants did not recognize the path split during Stage 1 for what it was, and thus eliminates any expectation that these participants could draw an inference from the fact that the AutoTurk used a name from one of the two paths. Another 6% of participants gave responses that indicated they believed their partner knew creatures from later in the game but not creatures from earlier in the game: this also indicates a lack of attention to (or memory for) the overall structure of the game. These participants were excluded from our analyses.

We first compared participants' judgments regarding AutoTurk's knowledge collected prior to Stage 2 to those collected in the Post-Test. If participants were behaving optimally, then we should expect that when interacting with *RedExpert*, participants' judgments should shift toward Yes for *Trimmelor* and *Narpelor* and toward No for *Molgirep* and *Bampirep*. When interacting with *BlueExpert*, participants' judgments should show the opposite pattern. And finally, when interacting with *EarlyFailure*, participants should show shifts toward No for all of the late-stage creatures. These patterns are indeed present in the data. We used a linear mixed effects regression model to predict whether the change from Stage 1 to Stage 2 judgments would be positive for individual creatures, with the AutoTurk's knowledge status as a fixed effect and the path chosen by the participant as a random effect. For *Trimmelor* (a Red Path creature), we found that participants interacting with *BlueExpert* were significantly less likely to have a positive shift ($\beta = -1.5$, S.E. = 0.4, $p < 0.001$), as were participants interacting with *EarlyFailure* ($\beta = -0.41$, S.E = 0.17, $p < 0.05$); participants interacting with *RedExpert* were significantly more likely to have a positive shift ($\beta = 0.89$, S.E = 0.14, $p < 0.001$). For *Narpelor* (another Red Path creature), we found a similar pattern; participants interacting with *RedExpert* were significantly more likely to have a positive shift ($\beta = 0.81$, S.E. = 0.13, $p < 0.001$) while participants interacting with *EarlyFailure* were significantly less likely to have a positive shift ($\beta = -0.3$, S.E = 0.14, $p < 0.05$), as were participants interacting with *BlueExpert* ($\beta = -1.09$, S.E. = 0.29, $p < 0.0001$). Overall, participants judgments following Stage 2 do reflect the evidence provided by AutoTurk's utterances.

3.4.2. Referring Expression Choice

In order to look at the choice of referring expression in a meaningful way, it was necessary to restrict all analyses to only those individuals who actually knew the relevant name for that trial, and because most participants chose the higher-paying Red Path, we focus our analyses on Trials 7 and 9, in which the target creatures are the two late-stage Red Path creatures. As shown in **Table 5**, both of these trials followed references to Blue and Red path creatures, and the form of the referring expression used by AutoTurk for those creatures varied depending on AutoTurk's

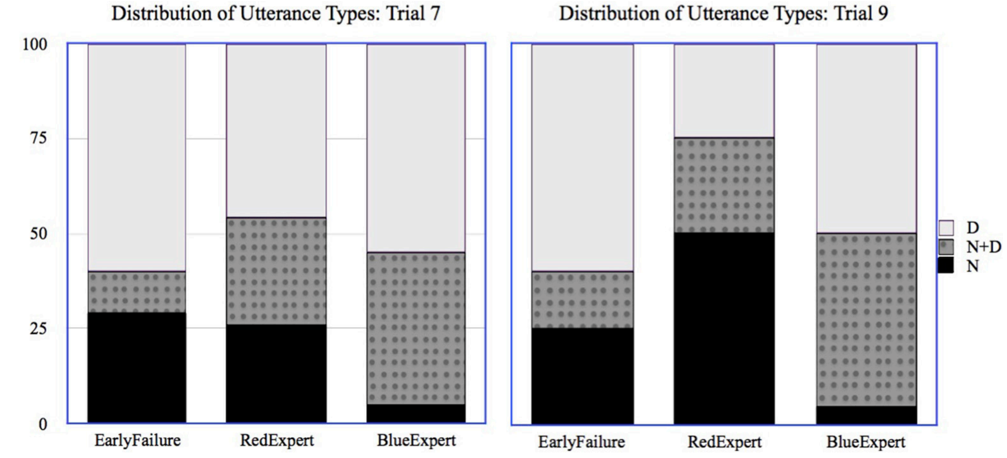


FIGURE 5 | Distribution of participants' utterance types when interacting with EarlyFailure, RedExpert, or BlueExpert for Trial 7 and Trial 9 (referring to Red Path creature.)

knowledge. **Figure 5** shows the distribution of utterance types chosen by participants for *Trimmelor* (Trial 7) and *Narpelor* (Trial 9).

For each trial, we used a mixed effects logistic regression model to predict whether the participant would use the N form for the target creature, with AutoTurk's knowledge as a fixed effect and the path chosen by the participant as a random effect. For Trial 7, we found that participants interacting with *BlueExpert* were significantly less likely to use the N form ($\beta = -2.39$, S.E. $= 1.15$, $p < 0.01$). However, we did not find any significant difference in the likelihood of using the N form between participants interacting with *RedExpert* and *EarlyFailure*; even when participants interact *EarlyFailure*, they appear to be as likely to use the N form as participants who interacted with RedExpert, who has already used a name for a (later) Red Path creature. Thus, the results of Trial 7 only partially support the hypothesis that participants are sensitive to their partners' use of name and adapt their choice of referring expression accordingly. Note, however, that *EarlyFailure* simply uses descriptions all of the time, and that this is not solid evidence of a lack of name knowledge the way that using a name for a creature from the other path is.

For Trial 9, we found that participants interacting with *RedExpert* were significantly more likely to choose the N form ($\beta = 2.7$, S.E. $= 1.2$, $p < 0.05$) than any others, and participants interacting with *BlueExpert* were significantly less likely to choose the N form ($\beta = -2.6$, S.E. $= 1.3$, $p < 0.05$). Thus, at least in the final trial of the experiment, and for an item whose name *RedExpert* had used in an N+D form in a previous trial, participants do seem to take AutoTurk's knowledge into account in their choice to use a name.

We also examined participants' Post-Test descriptions of their own strategy and their beliefs regarding their partner's strategy. Here, we found that most participants focused on the memory-related challenges of the task, commenting on how they kept the creatures' names straight and on how difficult that was. But based on the participants' *post-hoc* reflections on strategy, some individuals believed themselves to be sensitive to the information shared by their partner in the form of their choice of referring expression, and were aiming to make "optimal" referring expression choices based on that information. Many of these participants commented that they thought their partner was using the same strategy, but "doing a better job of it." But as we noted, the primary focus of participants in their Post-Test comments was on the challenges posed by the memory task; this may suggest that the basis for the inferences we were interested in testing (the overall structure of the game and the names contained within it) was either not recognized or misremembered by some participants.

3.5. Experiment 2 Discussion

Participants were able to generate more accurate beliefs regarding what their partner knew following interaction with that partner. In the Post-Test, participants changed their beliefs in the expected direction given the knowledge displayed by AutoTurk; thus, even in this simplified game environment, participants are capable of using their partner's utterances to arrive at a more accurate set of beliefs regarding their partner's knowledge.

The participants' referring expression choices also support our generalization hypothesis: in the final trial, we found that participants were significantly more likely to use a name to refer to final Red Path creature if they were interacting with *RedExpert*. Preceding trials paint a somewhat messier picture regarding the relationship between AutoTurk's utterances and the participants' choice of referring expression, particularly in the case of participants interacting with *EarlyFailure*. However, it's worth remembering that in Experiment 1, participants' use of names did not fully reflect their knowledge; participants often used descriptions even when they did, in fact, know the names. In the context of Experiment 2, this means that it is not necessarily valid to make the inference that because *EarlyFailure* used a description, it must not know the name. But we still would have expected participants interacting with *EarlyFailure* to use the N+D form, rather than using names by themselves, given the lack of evidence *for* knowing names. Still, participants interacting with *BlueExpert* were the least likely to use names; it seems that participants could in fact generalize from *BlueExpert*'s use of the N+D form for a Blue Path creature, and infer that *BlueExpert* could not possibly know the name of a Red Path creature.

Given the extent to which the kinds of belief updating we are interested in would depend on both accurate mental representations of the structure of the knowledge domain and attention to the partner's utterances, the combination of Mechanical Turk and novel knowledge may not have been a good one. Though many psychological findings have been successfully replicated using Mechanical Turk (e.g., Munro et al., 2010; Crump et al., 2013), the Mechanical Turk platform by itself does nothing to promote attention to the task unless the creator of the HIT creates incentives in the form of bonuses for work that meets some kind of standard. But in developing the bonus scheme to motivate participants to attend to the task in Experiment 1, we might have been probing participants' gambling behavior, rather than their conversational behavior: would a participant be willing to risk losing 10 points for the chance of gaining 18? What if the point spread were different? And indeed, many participants commented on making this calculation as if it were a bet, when they described their strategy in the Post-Test.

Even more important, though, is the difficulty that participants had remembering the names and the overall structure of the game in Stage 1. Based on their Pre-Test responses, participants struggled to remember which creature was which, and many of them seemed not to remember the order in which creatures were learned (or even that the path split meant that some creatures could not be learned together), and thus it seems unlikely that the kinds of memory associations that would be necessary in order for category-related cues to be useful for referring expression choice would even be present for these participants. These memory and knowledge-structure issues are crucial areas for future work; it seems likely, for example, that a sleep interval may be necessary in order to develop the kinds of concept associations necessary for these inferences (Stickgold and Walker, 2013; Landmann et al., 2014).

4. GENERAL DISCUSSION

Experiments 1 and 2 demonstrate that interlocutors are capable of using information gained over the course of conversation, particularly the information conveyed by the partner's choice of referring expression, to update their expectations regarding what knowledge is shared with their partner, and that these belief updates influence speakers' choice of referring expressions during subsequent conversation. In Experiment 2, we found some evidence for generalization, in that participants interacting with *BlueExpert* were the least likely to use the N form for a subsequent reference to a Red Path creature; the demonstrated knowledge of Blue Path names allowed participants to generalize to a necessary *lack* of knowledge of Red Path names. These findings have important implications for theories of CG use during conversation.

In many theories, the distinction between knowing that *a referent* is in CG (and is thus something to which the speaker could felicitously refer) and knowing that *a particular means of referring to that referent* is likely to be understood by the addressee, seems to be either blurred or non-existent. Yet even in those cases where an object is clearly in CG via a cue like visual co-presence (as in "cubbyhole" studies, e.g., Keysar et al., 2000), the speaker still has to decide what to call it. Even for common nouns, we usually have a choice regarding which to use to refer to a particular item: do we call something a *cassette* or a *tape*? While a number of factors influence this decision (word frequency, the other items in the display, etc), one of these should surely be whether or not the addressee is familiar with the link between a particular expression and the referent. This is especially true for proper names, which are arbitrary labels for a referent, and can only be understood by those who know about the link between the label and the referent. But it seems possible that various kinds of referring expressions could come to serve, under certain circumstances, as context-dependent conventions for referring (in other words, as context-dependent names), and this may be a way of connecting the work presented here to studies of lexical precedents or conceptual pacts (e.g., Metzing and Brennan, 2003; Brown-Schmidt, 2009). Partner-specific expectations relating to the means of referring could also play a role in comprehension, and in pilot work building off Wolter et al. (2011), we are currently exploring whether a speaker's use of a scalar contrast (e.g., *the big candle*) generates an expectation that the contrast item will be called *the small candle* even when the contrast is no longer present, by looking for reduced cohort competition effects with a competitor like *candy*.

We suggest that what is truly needed in an account of CG for definite reference is not triple co-presence of the sort described by Clark and Marshall (1978, 1981), but rather, quadruple co-presence: there must be some record in memory that links the speaker, the addressee, the object, and *a particular means of referring to that object* in order for a speaker to have a reasonable expectations that their use of that means of referring to that object will be understood by their addressee. Community membership can provide a powerful cue to such co-presence; we can often safely assume that certain means of referring to objects will be known by people by virtue of who they are and the communities to which they belong. This does not necessarily require any elaborate representations along the lines of the reference diaries proposed by Clark and Marshall; indeed, if this kind of knowledge can be drawn upon via the associative mechanisms proposed by Horton and Gerrig (2005a,b), it could underlie the kinds of belief-updating we described earlier. If we learn through conversation that our partner is the member of a particular community, this may trigger associations with the kinds of knowledge members of that community have (both in terms of referents and in terms of means of referring to those referents) that could lead to more accurate estimates of CG. And in particular, the evidence that could trigger such associations could come through our partner's use of an N+D form, as suggested by the results of Experiments 1 and 2, since that form enables speakers to display knowledge of a name without making any assumptions about whether it is shared with their addressee.

But not all displays of knowledge are equivalent. Finding out that an interlocutor knows the name *Statue of Liberty* or *guitar* is not particularly informative, as nearly any speaker of English would know those names. But evidence that the speaker knows the name *South Street Seaport* or *viol da gamba* should cause their conversational partner to shift toward believing their fellow interlocutor has expertise with New York City landmarks or Renaissance-era stringed instruments, respectively. If language comprehension is, in part, a process of trying to explain why the speaker said what they said the way that they said it (as in, e.g., Hobbs et al., 1993), then the explanation we seek may be rooted in what we think the speaker knows. Bayesian belief-updating could provide a useful framework for exploring the extent to which interlocutors use evidence to generate such explanations in a rational way, based on the informativity of the evidence provided by their partner. We have ample evidence that during interactive conversation, interlocutors adapt their expectations regarding such things as the likelihood of particular syntactic constructions (e.g., Kleinschmidt et al., 2012; Fine et al., 2013) or of particular phonetic realizations (e.g., Kleinschmidt and Jaeger, 2015), and both of these adaptation phenomena have been successfully explored using Bayesian belief-updating models. Future work in this area should focus on understanding the process by which people generate, update, and generalize beyond their prior expectations regarding partner knowledge; games like the ones used in Experiments 1 and 2 may provide a fruitful paradigm for exploring these processes in more detail, particularly as they relate to the relative informativity of a particular knowledge-display.

The choice of referring expression is a valuable tool for exploring questions about how we use CG. Even in those cases where an object is physically co-present and therefore can be referred to using a definite NP, speakers still must decide what to call it; thus, it is crucial not to think about CG simply in terms of what interlocutors know about what their conversational partner can *see*, but to also consider what interlocutors believe about what their partner *knows*. What's in a name? Evidence about knowledge.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

We would like to acknowledge Chelsea Marsh for her assistance with recruiting, training, scheduling, and testing participants in Experiment 1, and members of the *KurTan* lab and the Center for Language Sciences at the University of Rochester for fruitful discussion and feedback. *Funding:* This research was supported by National Institute of Child Health and Human Development Grant HD-27206 to MT.

REFERENCES

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London; New York, NY: Routledge.

Birch, S., and Bloom, P. (2007). The curse of knowledge in reasoning about false belief. *Psychol. Sci.* 18, 382–386. doi: 10.1111/j.1467-9280.2007.01909.x

Brennan, S., and Hanna, J. (2009). Partner-specific adapation in dialogue. *Top. Cogn. Sci.* 1, 274–291. doi: 10.1111/j.1756-8765.2009.01019.x

Brown-Schmidt, S. (2009). Partner-specific interpretations of maintained referential precedents during interactive dialog. *J. Mem. Lang.* 61, 171–190. doi: 10.1016/j.jml.2009.04.003

Brown-Schmidt, S. (2012). Beyond common and privileged: gradient representations of common ground in real-time langauge use. *Lang. Cogn. Process.* 27, 62–89. doi: 10.1080/01690965.2010.543363

Brown-Schmidt, S., Gunlogson, C., and Tanenhaus, M. (2008). Addresses distinguish shared from private information when interpreting questions during interactive dialog. *Cognition* 107, 1122–1134. doi: 10.1016/j.cognition.2007.11.005

Brown-Schmidt, S., and Hanna, J. E. (2011). Talking in another person's shoes: incremental perspective-taking in language processing. *Dialogue Discourse* 2, 11–33. doi: 10.5087/dad.2011.102

Brown-Schmidt, S., and Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted langauge game approach. *Cogn. Sci.* 32, 643–684. doi: 10.1080/03640210802066816

Clark, H., and Krych, M. (2004). Speaking while monitoring addressees for understanding. *J. Mem. Lang.* 50, 62–81. doi: 10.1016/j.jml.2003.08.004

Clark, H., and Marshall, C. (1981). "Definite reference and mutual knowledge," in *Elements of Discourse Understanding*, eds A. K. Joshi, B. Webber, and I. Sag (Cambridge: Cambridge University Press), 10–63.

Clark, H. H., and Marshall, C. R. (1978). "Reference diaries," in *Theoretical Issues in Natural Language Processing*, Vol. 2, ed D. L. Waltz (New York, NY: Association for Computing Machinery), 57–63.

Creel, S. C., Aslin, R. N., and Tanenhaus, M. K. (2008). Heeding the voice of experience: the role of talker variation in lexical access. *Cognition* 106, 633–664. doi: 10.1016/j.cognition.2007.03.013

Creel, S. C., and Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *J. Mem. Lang.* 65, 264–285. doi: 10.1016/j.jml.2011.06.005

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE* 8:e57410. doi: 10.1371/journal.pone.0057410

Epley, N., Van Boven, L., Keysar, B., and Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *J. Pers. Soc. Psychol.* 87, 327–339. doi: 10.1037/0022-3514.87.3.327

Fine, A. B., Jaeger, T. F., Farmer, T. A., and Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE* 8:e77661. doi: 10.1371/journal.pone.0077661

Fussell, S., and Krauss, R. (1992). Coordination of knowledge in communication: effects of speaker' asssumptions about what others know. *J. Pers. Soc. Psychol.* 62, 378–391. doi: 10.1037/0022-3514.62.3.378

Galati, A., and Brennan, S. E. (2010). Attenuating repeated information: for the speaker, or for the addressee? *J. Mem. Lang.* 62, 35–51. doi: 10.1016/j.jml.2009.09.002

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychol. Rev.* 105, 251–279. doi: 10.1037/0033-295X.105.2.251

Gorman, K. S., Gegg-Harrison, W. M., Marshall, C. R., and Tanenhaus, M. K. (2013). What's learned together stays together: speakers' choice of referring expression reflects shared experience. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 843–854. doi: 10.1037/a0029467

Grosz, B., and Sidner, C. (1986). Attention, intention, and the structure of discourse. *Comput. Linguist.* 12, 175–204.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307.

Heller, D., Gorman, K. S., and Tanenhaus, M. K. (2012). To name or to describe: shared knowledge affects referential form. *Top. Cogn. Sci.* 4, 290–305. doi: 10.1111/j.1756-8765.2012.01182.x

Hobbs, J. R., Stickel, M. E., Appelt, D. E., and Martin, P. (1993). Interpretation as abduction. *Artif. Intell.* 63, 69–142. doi: 10.1016/0004-3702(93)90015-4

Horton, W. S., and Gerrig, R. (2005a). Conversational common ground and memory processes in language production. *Discourse Process.* 20, 1–35. doi: 10.1207/s15326950dp4001_1

Horton, W. S., and Gerrig, R. (2005b). The impact of memory demands on audience design during language production. *Cognition* 96, 127–142. doi: 10.1016/j.cognition.2004.07.001

Isaacs, E. A., and Clark, H. H. (1987). References in conversations between experts and novices. *J. Exp. Psychol. Gen.* 116, 26–37. doi: 10.1037/0096-3445.116.1.26

Keysar, B., Barr, D., Balin, J., and Brauner, J. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychol. Sci.* 11, 32–38. doi: 10.1111/1467-9280.00211

Kleinschmidt, D. F., Fine, A. B., and Jaeger, T. F. (2012). "A belief-updating model of adaptation and cue combination in syntactic comprehension," in *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (Sapporo), 605–610.

Kleinschmidt, D. F., and Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* 122, 148–203. doi: 10.1037/a0038695

Kronmüller, E., and Barr, D. (2015). Referential precedents in spoken language comprehension: a review and meta-analysis. *J. Mem. Lang.* 83, 1–19. doi: 10.1016/j.jml.2015.03.008

Lambrecht, K. (1994). *Information Structure and Sentence Form: Topic, Focus, and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Landmann, N., Kuhn, M., Piosczyk, H., Feige, B., Baglioni, C., Speigelhalder, et al. (2014). The reorganization of memory during sleep. *Sleep Med. Rev.* 18, 531–541. doi: 10.1016/j.smrv.2014.03.005

Metzing, C., and Brennan, S. (2003). When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *J. Mem. Lang.* 49, 201–213. doi: 10.1016/S0749-596X(03)00028-7

Munro, R., Bethard, S., Kuperman, V., Lai, V., Melnick, R., Potts, C., et al. (2010). "Crowdsourcing and language studies: the new generation of linguistic data," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Stroudsburg, PA), 122–130.

Nadig, A., and Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* 13, 329–336. doi: 10.1111/j.0956-7976.2002.00460.x

Roberts, C. (2004). Uniqueness in definite noun phrases. *Linguist. Philos.* 26, 287–350. doi: 10.1023/A:1024157132393

Stickgold, R., and Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nat. Neurosci.* 16, 139–145. doi: 10.1038/nn.3303

Tanenhaus, M., and Brown-Schmidt, S. (2008). Language processing in the natural world. *Philos. Trans. R. Soc. B Biol. Sci.* 363, 1105–1122. doi: 10.1098/rstb.2007.2162

Wolter, L., Gorman, K. S., and Tanenhaus, M. K. (2011). Scalar reference, contrast, and discourse: separating effects of linguistic discourse from availability of the referent. *J. Mem. Lang.* 65, 299–317. doi: 10.1016/j.jml.2011.04.010

Wu, S., and Keysar, B. (2007). The effect of information overalp on communication effectiveness. *Cogn. Sci.* 31, 1–13. doi: 10.1080/03640210709336989

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The Role of Metarepresentation in the Production and Resolution of Referring Expressions

William S. Horton[1]* and *Susan E. Brennan*[2]

[1] Department of Psychology, Northwestern University, Evanston, IL, USA, [2] Department of Psychology, Stony Brook University, Stony Brook, NY, USA

In this paper we consider the potential role of metarepresentation—the representation of another representation, or as commonly considered within cognitive science, the mental representation of another individual's knowledge and beliefs—in mediating definite reference and common ground in conversation. Using dialogues from a referential communication study in which speakers conversed in succession with two different addressees, we highlight ways in which interlocutors work together to successfully refer to objects, and achieve shared conceptualizations. We briefly review accounts of how such shared conceptualizations could be represented in memory, from simple associations between label and referent, to "triple co-presence" representations that track interlocutors in an episode of referring, to more elaborate metarepresentations that invoke theory of mind, mutual knowledge, or a model of a conversational partner. We consider how some forms of metarepresentation, once created and activated, could account for definite reference in conversation by appealing to ordinary processes in memory. We conclude that any representations that capture information about others' perspectives are likely to be relatively simple and subject to the same kinds of constraints on attention and memory that influence other kinds of cognitive representations.

Keywords: reference, common ground, metarepresentation, memory, cognitive models

OPEN ACCESS

Edited by:
Kees Van Deemter,
University of Aberdeen, UK

Reviewed by:
Stanka A. Fitneva,
Queen's University, Canada
Dale J. Barr,
University of Glasgow, UK

****Correspondence:***
William S. Horton
whorton@northwestern.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *14 September 2015*
Accepted: *11 July 2016*
Published: *27 July 2016*

Citation:
Horton WS and Brennan SE (2016) The Role of Metarepresentation in the Production and Resolution of Referring Expressions. Front. Psychol. 7:1111. doi: 10.3389/fpsyg.2016.01111

REFERRING AND REPRESENTATION

Speakers have many options in the production of referring expressions, ranging from simple pronouns to complex definite or indefinite noun phrases. Moreover, there is potential for substantial variability in the noun phrases speakers choose. Consider just a few of the referring expressions for the novel object in **Figure 1**, each used by a different pair of speakers.

Successful referring—with the desired effect of getting a speaker and an addressee focused on the same referent—is as much about an interactive *process* as it is about a *product*. The expressions in **Figure 1** emerged from dialogues such as the following, collected during an experiment by Stellmann and Brennan (1993), in which a director, A, is trying to help a matcher seated behind a screen, B, to match a dozen cards with abstract geometric figures (tangrams) into a target order (Note: Overlapping speech occurs within *asterisks*).

A: Number 8... is a candle?
B: A candle... ok...
A: It's got a lot*of*
B: *is it* wider on top and short on the bottom and it has, like, a diamond sticking out from the top?

"a bat"
"the candle"
"the anchor"
"the rocket ship"
"the Olympic torch"
"the Canada symbol"
"the symmetrical one"
"shapes on top of shapes"
"the one with all the shapes"
"the bird diving straight down"
"the airplane flying straight down"
"the angel upside down with sleeves"
"the man jumping in the air with bell bottoms on"

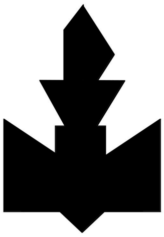


FIGURE 1 | Stellmann and Brennan (1993), unpublished data.

A: Yes but rotate it the other way and so it's wider on the bottom
B: Wider on the bottom, hold on...
A: If you um for instance it has um
B: Which one is um number 8
A: This is number 8 if you want to think of it also it looks like someone doing a split on the ground or jumping in the air, yeah, with bell bottoms on
B: Which direction?
A: *uh*
B: *oh* alright, alright, I see it, I see it
A: Ok
B: It looks like it's doing a split in the air
A: Right *exactly*
B: *alright*

Here, A begins by proposing that the object in question resembles a candle, but B is unable to recognize any suitable object in the set and asks for clarification. A, guessing which object B might be considering, suggests rotating the card. B attempts to discuss the card's geometric details, but after a few exchanges focused on the geometry of the object, A tosses out a new counter-proposal (*someone doing a split on the ground or jumping in the air, yeah, with bell bottoms on*). B eventually confirms this perspective and places the corresponding card on the board; this action provides evidence for the success of the referring process. In this exchange, after 16 conversational turns, speaker and addressee have finally come to believe that they mean the same thing. Thereafter, A and B referred to this object again on subsequent rounds as follows (note that the exchanges in these rounds are each separated by matches to an average of 11 other objects):

Round 2: B: The person with the bell bottoms doing a split in the air
A: Ok um wait alright ok
Round 3: A: The person with the bell bottoms jumping in the air
B: Got it
Round 4: B: The bell bottomed jumper
A: I had it there man

This example is representative of how 26 pairs of subjects in this experiment referred to the 12 objects that they repeatedly matched over four rounds. The processes through which people interactively seek and provide evidence for shared perspectives during referring in conversation is known as *grounding* (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Clark and Brennan, 1991). The consistency among referring expressions produced by A and B across rounds, as well as the evidence provided in each round by the successful match of the matcher's object to the director's, suggest that for each object, the two partners built up common ground that enabled them to mentally represent the object in the same way (or highly similar ways). As the same time, the variability of expressions in **Figure 1** suggests that other pairs conceptualized this object quite differently. Through such processes of conceptual coordination, partners converge on, and re-use the same terms within a conversation, displaying lexical and conceptual *entrainment* in their choices of referring expressions (Garrod and Anderson, 1987; Brennan and Clark, 1996).

What sorts of cognitive mechanisms support such coordination? There have been several types of answers to this question. Perhaps the simplest is one that appeals to direct cross-speaker activation of particular expressions, as proposed by the *interactive alignment* account (Pickering and Garrod, 2004). This account presumes that interlocutors converge on the same referring expressions simply because one speaker's utterances can automatically prime similar responses from the other, facilitating similar discourse representations over time (assuming that the interlocutors are similar). A compatible view, from Brown and Dell (1987; Dell and Brown, 1991), suggests that what appears to be partner-specific coordination is often actually *generic*, in that what is easy for a speaker to produce tends to be easy for an addressee to understand. Relevant claims for this view include that "speakers and listeners do not routinely take common ground into account during initial processing" (Pickering and Garrod, 2004, 179), and that "normal conversation does not routinely require modeling the interlocutor's mind" (ibid, p. 180). Pickering and Garrod's proposal is consistent with models that propose that common ground is used only *on demand*, when repair is needed (e.g., Brown and Dell, 1987; Horton and Keysar, 1996; Keysar et al., 2000):

> "Establishment of full common ground is, we argue, a specialized and non-automatic process that is used primarily in times of difficulty (when radical misalignment becomes apparent)... speakers and listeners do not routinely take common ground into account during initial processing... full common ground is only used when simpler mechanisms are ineffective" (Pickering and Garrod, 2004, p. 179).

Much experimental work by Keysar, Barr and colleagues (e.g., Keysar et al., 2000; Barr and Keysar, 2002; Keysar et al., 2003) has been presented in support of this idea, suggesting that interlocutors in conversation behave egocentrically (at least at first), but that this does not hamper communication as long as interlocutors are similar enough and happen to inhabit the same context (see also Shintel and Keysar, 2009).

Another type of answer to the question of conversational coordination involves the notion of *metarepresentation*, which generally refers to the representation of another representation. Sperber (2000, p. 3) identified four main categories of metarepresentation: "Mental representations of mental representations (e.g., the thought, "John believes that it will rain"), mental representations of public representations (e.g., the thought, "John said that it will rain"), public representations of mental representations (e.g., the utterance, "John believes that it will rain"), and public representations of public representations (e.g., the utterance, "John said that it will rain")." Accounts of language use often assume, either implicitly or explicitly, that definite reference requires some form of the first of these, or the mental representation of another's *mental* representation—typically considered as the representation of another person's knowledge, needs, or beliefs. Certainly, Grice's (1975) original notion of conversational implicature was inherently metarepresentational in this sense, being rooted in the idea that pragmatic meanings involve direct consideration of what is mutually known between rational speakers. Similarly, linguistic theories of reference production frequently assume that choices in the form and content of referring expressions emerge from speaker's assessments of the accessibility of particular referents in the minds of their addressees (e.g., Gundel et al., 1993; Grosz et al., 1995). And models of communication and intention recognition in computational linguistics and logic have often been focused on providing formalizations of representations of other agent's epistemic states (e.g., Cohen and Perrault, 1979; Ditmarsch et al., 2007). An important question for all of these approaches, of course, is how well they succeed in capturing the kinds of knowledge state inferences that human interlocutors are likely to make in real time during genuine interactions.

Theories appealing to the mental representation of other's mental representations often presume detailed consideration of the needs, knowledge, or beliefs of a social partner. For example, the influential notion of *theory of mind*, as first articulated by Premack and Woodruff (1978), and refined further by Dennett (1978) and Pylyshyn (1978), refers to an individual's mental capacity to reason about the mental states of others, an ability that appears to follow a distinct developmental trajectory into adulthood (Wellman et al., 2001; Apperly, 2011, 2013). An even more complex way to think about metarepresentation invokes the recursive modeling of mutual knowledge (*I know X; you know X; I know that you know X; you know that I know X; I know that you know that I know X*; and so on); however, it is widely acknowledged that this sort of recursive reasoning would be so resource-intensive as to be implausible [see the debates in Smith (1982) on this *mutual knowledge paradox,* and Clark and Marshall's (1978, 1981) proposed solution involving inferences about co-presence].

Given the apparent ease with which interlocutors plan and resolve referring expressions in conversation, it might seem prudent to prefer the simplest account that relies on "priming" of referent-label associations across interlocutors. However, experimental corpora such as Stellmann and Brennan (1993) raise some key questions about the nature of the representations underlying referential communication that cannot be explained by simple associations alone (see similar evidence presented by Brennan and Clark, 1996 and Horton and Gerrig, 2005b).

To that end, it is important to note that Stellmann and Brennan's experiment actually involved quartets of speakers. Two additional subjects, C and D, matched the same cards at the same time as A and B, but in a neighboring room. For the item shown in **Figure 1**, C and D entrained on a perspective that they ended up labeling as *the anchor*. Crucially, after both pairs matched the cards for Rounds 1–4, they were split up and re-paired such that A and D completed Rounds 5–8 together, as did B and C. Of interest was whether there would be any savings in the linguistic effort needed to match these now-familiar objects. Here is what ensued between A and her new partner, D, in Round 5 immediately after the partner switch:

A: ah the second one looks like maybe a person jumping in the air who is wearing bell bottoms... or it could be a candle
D: Jumping with bell bottoms...
A: yeah... oh well, you know, cause it has two triangles coming from the left and the right, but they're, um, it looks like a person jumping... he's not- it's definitely symmetric down the middle...
D: oh man...
A: There's- and it's a- um, a diamond, a triangle, a rectangle, and two, ah, two triangles going from the left and right, you know, you could, ah...
D: You can't make a picture out of it?
A: ah let's see, if you put it on its side it looks like an E
D: an E?
A: yeah, no, let's call it an anchor ∗that's cool∗
D: ok *the anchor one* yeah ok
A: ok it looks like an anchor
D: *yeah*
A: *that's cool*

What is striking is that the director, A, did not simply pick up where had she left off with B, with the concise expression that had worked most recently ("the bell bottomed jumper")—even though this would presumably have corresponded to the strongest trace in memory (according to Garrod and Anderson, 1987 *output-input coordination principle*, a precursor to the interactive alignment theory). Instead, she proposed an indefinite referring expression marked as tentative by hedging ("*ah the second one looks like maybe* a person jumping in the air who is wearing bell bottoms"), as if to display sensitivity to the fact (that is, mentally representing) that she as yet had no common ground with D. She also proffers an alternative expression, "or it could be a candle." Such *re-conceptualizations* have been observed in other referential communication studies that involve switching from an old to a new conversational partner (e.g., Brennan and Clark, 1996; Horton and Gerrig, 2005b; Gorman et al., 2013). After D failed to accept either of the (re-conceptualized) perspectives that A had discussed with her old partner, the new partners ended up converging on the perspective that D happened to entrain on earlier with C in the other room. Meanwhile, in that other room, B and C struggled valiantly (over 27 turns) to arrive at what might best be described as a hybrid perspective (*bell-bottom anchor*), which they continued to use in their next 3 rounds together. Although all of these people were, by then, individually quite familiar with the object in **Figure 1**, they still had to expend significant effort to ground their references to this object with their new partners in Round 5, just as with their initial partners in Round 1.

In the last stage of Stellmann and Brennan's experiment, the original partners were reunited for four final rounds, 9–12: A joined up again with B again, and C with D, whereupon they matched the same cards again. At that point, A and B reverted immediately and efficiently to the perspective they had entrained on previously, using "the guy with the bell bottoms jumping," "the bell bottom," "the bell bottomed guy," and "the bell bottomed man" respectively in Rounds 9–12. Likewise, C and D immediately returned to the unadorned definite expression "the anchor" as soon as they got back together in Round 9. That pairs often switched back smoothly in Round 9 to the relatively short expression they had entrained upon in rounds 1–4 suggests that their representations included more than just the association of a referent and a referring expression (or the perspective it indexes), but information about the communication partner as well.

In this article, we suggest that simple associations are not sufficient to account for these and similar patterns of conversational referring. As an alternative possibility, we examine the role of metarepresentations in communication, and whether representations of a partner's goals, informational needs, or knowledge must, by necessity, involve the kinds of time-consuming inferences most commonly associated with theory of mind. We consider whether metarepresentations are created, maintained, and used routinely during conversational episodes, or only strategically in response to special circumstances such as evidence of misunderstanding. At the same time, we will appeal to an explanation relying on ordinary memory processes such as resonance (Ratcliff, 1978), and avoid positing any sort of "special" memory representation or separate stage of processing to account for the apparent effects of common ground upon referring in conversation. We conclude that a simple (meta) representation about a conversational episode (once created and activated) could be rapidly re-instantiated into the current discourse context via simple cues.

METAREPRESENTATION IN THEORY OF MIND

As stated previously, the notion of metarepresentation has been especially important within the literature on theory of mind (ToM), which refers to the capacity to reason about the mental states of others. A large body of research has sought to answer such questions as whether theory of mind is a uniquely human capacity, whether deficits in theory of mind can help explain particular social and communicative disorders such as autism and schizophrenia, and whether theory of mind abilities are mediated by unitary, specialized neural circuits in the brain (e.g., Call and Tomasello, 2008; Frith and Frith, 2012; Baron-Cohen et al., 2013).

As a rule, ToM is fundamentally important for making sense of the social world. To give a simple example, imagine you observe someone pick up a key and walk with it toward a closed door. Based on your capacity for mental state attribution (and your knowledge of keys and doors), you might reasonably infer that this person has the intention of unlocking the door. Some (e.g., Perner and Ruffman, 2005; Penn and Povinelli, 2007) have argued that expectations about another's behavior could be generated primarily on the basis of associative or statistical knowledge concerning the kinds of actions that involve, for example, keys and doors, without being mediated by a theory of mind inference, although others have argued that simple associations or rules cannot account for the range of contexts across development in which young children, in particular, come to show evidence for perspective-taking (Baillargeon et al., 2010). Because a person's (private) intentions are not directly observable, one can only use observations of behavior to make inferences about the mental states giving rise to those behaviors. Making such inferences is a form of "mindreading"—reasoning about another's internal mental state—and is the hallmark of how people engage their theory of mind (Apperly, 2011).

In principle, mindreading can involve a wide range of mental state attributions, including inferences about emotional states, perceptual access, and desires and goals. The empirical literature on ToM (and certainly the developmental literature) has commonly focused on situations that involve deliberative, reflective assessments of another person's knowledge, most often involving "false belief," in which another's knowledge conflicts with reality (Baron-Cohen et al., 1985; Wellman et al., 2001). For example, the classic Sally-Ann task (Wimmer and Perner, 1983) asks children to reason about Sally's belief concerning the location of an object, which, unbeknownst to her, has been moved from its original location. To pass this task, a child must recognize that Sally possesses an incorrect belief about the object's location, be able to understand that the question is asking where she *would* (rather than *should*) look for it, and respond accordingly to the experimenter's question. Quite clearly, this requires access to metarepresentations of Sally's beliefs. One must not only represent what Sally believes, but must also be able to appreciate how Sally's beliefs differ from one's own (and from reality). The representational and computational challenges involved in such situations have been cited as one reason why children don't consistently pass classic false belief tests until after the age of four (e.g., Call and Tomasello, 1999; Bloom and German, 2000).

Much of the literature on theory of mind, though, suggests that adults, at least, have the ability to construct and access representations of another's knowledge, even if they don't always apply this ability in situations that require puzzling out what another person is likely to believe, or how they are likely to act (Keysar et al., 2003). It is not immediately evident, though, whether the types of deliberative mental state attribution required by experimental tasks exploring false belief and theory of mind are qualitatively similar to the kinds of spontaneous inferencing regarding common ground that would seem to take place during routine conversation. Moreover, the types of mental state attribution commonly presumed within the literature on theory of mind have often assumed a view of metarepresentations as discrete, neatly packaged, deterministic bits of information about other people's knowledge and beliefs.

Experiments exploring this capacity most often probe for "on-demand" inferences based on the presence or absence of knowledge, as in the classic Sally-Ann task, rather than measuring probabalistic inferences that occur spontaneously. Exceptions include work by Samson, Apperly and colleagues (e.g., Samson et al., 2010; Surtees and Apperly, 2012; see also van der Wel et al., 2014) who have shown that both children and adults can use extremely simple cues, such as the direction of attention, to rapidly generate inferences about mental states. Moreover, responding correctly to the classic Sally-Ann task requires the ability to respond to a question that depends on understanding modal verbs. When 3-year-olds are simply asked to act out "what happens next?" upon Sally's return to the hidden-object situation, they are more likely to demonstrate a spontaneous ability to reason about ToM and respond correctly (Rubio-Fernández and Geurts, 2013).

Such evidence suggests that, while it can take time to reason (on demand) about another person's knowledge or beliefs, once a relevant metarepresentation has been evoked, taking account of a partner context need not involve a time-consuming (or even a very mature) process of reasoning. This information can be used just like any other information in memory.

REFERENCE DIARIES AND TRIPLE CO-PRESENCE

Metarepresentation in some form is often invoked by accounts of referential communication. Perhaps the most influential account of definite reference comes from Clark and Marshall (1978, 1981), who observed, "people's memory must be organized to enable them to get access to evidence they will need to make felicitous references. To make or interpret definite references people have to assess certain "shared" knowledge. This knowledge, it turns out, is defined by an infinite number of conditions. How then can people assess this knowledge in a finite amount of time?" (Clark and Marshall, 1981, pp. 56–57). After rejecting recursively-achieved mutual knowledge as cognitively implausible, Clark and Marshall proposed that interlocutors take advantage of representations they called "reference diaries" that encode evidence for *triple co-presence*—or evidence that the speaker, addressee, and referent were "openly co-present together" (1981, p. 32). On this account, definite expressions (such as those in **Figure 1**) are constructed and interpreted against the common ground established by interlocutors through a heuristic that shortcuts the problem of computing mutual knowledge recursively. This heuristic is based on an inference that the parties in a conversation perceive or recall common ground based on what they've discussed together (linguistic co-presence), their experiences together in the same environment (physical co-presence), or their presumed socio-cultural overlap (community co-membership). *Prior* co-presence is established through previous experience together, whereas *potential* co-presence is evoked by a speaker's rational expectation that an addressee can use the current context to understand, for example, the intended referent of *I'd like that loaf of bread please* when accompanied by a pointing gesture over the addressee's shoulder (Clark and Marshall, 1981; for discussion, see Polichak and Gerrig, 1998). An inference on the part of a speaker that she and her addressee are contextually co-present presumably supports some kind of suitable representation of partner-specific information that facilitates *audience design* (Clark and Murphy, 1982; Bell, 1984; Horton and Gerrig, 2002), allowing her to produce referring expressions that a particular addressee is likely to be able to resolve. Such representations also allow addressees to interpret the same referring expression differently when it is spoken by different speakers in different contexts (Metzing and Brennan, 2003).

Clark and Marshall (1981) were not specific about the characteristics or limitations of possible partner-specific representations (which we call metarepresentations), apart from proposing that they encode triple co-presence (an association linking the self, an other, and the information in question). However, in their sections on *organization of memory* and *components of memory*, they referred to episodic memories of events that speakers and addressees have experienced together as "compartmentalized into useful units" that can be selectively accessed (p. 55), and that shift when the interlocutor changes in a conversation.

Ongoing debates in psycholinguistics have focused on the extent to which consideration of common ground *routinely* and *initially* informs language processing (Brennan and Hanna, 2009), or whether it is invoked in a separate stage of processing, like a repair (e.g., Brown and Dell, 1987; Keysar et al., 2000; Pickering and Garrod, 2004) Where Clark and Marshall and their critics agree is that sometimes effort must be expended in order to establish common ground or to propose or resolve a referring expression, but that frequently, referring in conversation seems effortless. The question remains as to what sort of representation underlies processing in this latter situation.

A ROLE FOR METAREPRESENTATIONS IN REFERRING

In principle, the notion of metarepresentation would appear to be central to models of reference and language use. An important question is whether considering metarepresentations need always be resource-intensive and time-consuming, or whether people can rapidly, and potentially automatically, behave as if they have access to appropriate metarepresentations under the right circumstances. In this section and the next, we argue that the metarepresentations themselves need not be elaborate or encode inferences about complex mental states, but can be simple and partial, driven by current conversational purposes. Furthermore, we propose that once a suitable episodic metarepresentation has been activated, it may be used as fluidly and rapidly as any other information available in memory.

In contrast, some have argued that speakers and addressees are inevitably "egocentric," and that taking account of a partner's perspective as distinct from one's own can happen only as a kind of delayed processing or repair (Keysar et al., 1998; Epley et al., 2004). Shintel and Keysar (2009) point out that "elaborate reasoning that requires interlocutors to keep updated metarepresentations of the other's beliefs that are separate from their own representations of the situation is both time consuming and cognitively demanding." Not surprisingly, experiments that place speakers in perceptual contexts with both salient privileged information *and* information that is in common ground with an addressee, along with the need to continuously distinguish these, show evidence for interference between dueling perspectives (Horton and Keysar, 1996; Keysar et al., 2000). Indeed, some have used this kind of evidence to argue for modularity in cognitive architecture (Barr, 2008).

A related argument is that what appears to be partner-specific processing (resulting, e.g., in entrainment and audience design) occurs simply when speakers and hearers happen to share the same context, as suggested by Pickering and Garrod's (2004) interactive alignment model (see also Brown and Dell 1987). As Pickering and Garrod argue, simple priming can facilitate convergence in referring expressions without obligating interlocutors to directly represent the other person's perspective. On these accounts, metarepresentations would have little role to play in fundamental aspects of language processing, instead being relevant only in the context of slower, more effortful processes of monitoring and repair.

In this context, we make two critical points about referring in conversation. The first critical point is that referring is not a deterministic process, but a collaborative one that involves coordination between (at least) two people. As a result, any form of metarepresentation that emerges from conversational grounding is likely to be highly probabilistic, in the sense that the relevant memory traces will vary in strength and accessibility. To understand why, we return to our examples from Stellmann and Brennan's corpus and focus more closely on what happened after the partner-switch at Round 5. Here, D began with a couple of proposals to her new partner, B; she first proposed the perspective that had worked well earlier with her former partner C, and then added another perspective that C had failed to take up:

D: ah the second one looks like maybe a person jumping in the air who is wearing bell bottoms... or it could be a candle
B: Jumping with bell bottoms...

Here, by echoing some of D's words hesitantly, B provides evidence of uncertainty; he's considering this proposal but isn't able to accept it. D responds by trying to motivate *the jumping man with bell bottoms* using a lengthy and laborious geometric description, but then B suggests returning to a figurative strategy:

B: You can't make a picture out of it?

After two more proposals, D fortuitously hits upon the perspective that B happened to use earlier with former partner A: "let's call it an anchor." This works, and on they go. Examples like this underscore a critical aspect of Clark and Marshall's original co-presence account (Clark and Brennan, 1991; Clark, 1996), which is that heuristic-based approximations of what others know may suffice much of the time, given that grounding and the potential for interaction provide relatively inexpensive ways to recover *if* and *when* interlocutors get it wrong. Speakers' current purposes often don't require them to be perfect or to work too hard on the inferences they make. They are thus able to balance the costs of delaying an utterance (in order to plan it more fully) with the risks of appearing to be inattentive or losing their partners' attention or running out of time in a task (Clark and Brennan, 1991).

A second critical point is that evidence for representation can be found in *how* a speaker presents a particular referring expression. Referring expressions can be fluent, disfluent, brief, wordy, hedged, or presented with falling or question intonation (Smith and Clark, 1993; Brennan and Williams, 1995). As the examples from Stellmann and Brennan's corpus illustrate, a referring expression, once proffered, has the potential to be (and be marked as) tentative, vague, or unacceptable, suggesting that any representation of another's knowledge that might emerge through conversational grounding also has the potential to be incomplete or incorrect. Moreover, given that referring expressions and other types of utterances are generally produced and understood on a time scale that would seem to preclude lengthy deliberation, there must be processes that permit reasonably rapid access to perspective-relevant information. These factors impose critical constraints upon any account attempting to capture the nature of successful referring in language use.

As the speakers in Stellmann and Brennan's study transitioned from one partner to another and back again, the way in which they framed referring expressions changed accordingly, as described previously. One possible explanation for the hedging after the first partner change in Round 5 is that the presence of the brand-new partner, D, weakened the memory trace for the previous referring expression. However this explanation by itself is not so convincing, as the repeated referring in Rounds 1–4 with partner B should have rendered A's memory for what to call this object quite strong. Moreover, the return to Partner B in Round 9 (when the original pairs were reunited) showed little to no disruption due to the partner switch, but rather a smooth return to the originally entrained-upon perspective. Reverting to this original perspective might be expected to fight against one's memory traces for the repeated (and more recent) references to *the anchor* in Rounds 5–8. However, a plausible explanation is that the partner-switch successfully boosted accessibility of the previous episode, likely through a *compound cue* comprised of the current referent plus the presence of the original partner B (Ratcliff and McKoon, 1988; Horton and Gerrig, 2005a, in press).

At issue is when and how such a shift in referring expressions also might be shaped by remembering that the original partner shares a particular perspective on the tangrams not shared by the other partner. If so, we can ask if and when this belief is encoded as part of the original memory trace for this interaction. The tentative way in which A expresses "looks like maybe a person jumping in the air who is wearing bell bottoms" is presumably *not* due to A's own lack of facility with this perspective (which, after all, A had just finished deploying over and over with B), but may, in fact, be caused by A's sense that a new partner, D, would not find it so easy to take this perspective (in other words, that D is not implicated in A's episodic representation). In principle, the difference between proposing the efficiently packaged definite expression "bell bottomed jumper" vs. "looks like maybe a person jumping in the air who is wearing bell bottoms" would seem to involve ready access to suitable representations that encode such beliefs about another's knowledge. As we suggest, though, the availability of such representations is likely to be influenced by the nature of the grounding processes that give rise to these beliefs, providing speakers with the opportunity to encode inferences concerning what can be taken as shared—especially when these inferences are supported by salient features of the conversational context.

It is not clear how this apparently partner-specific effect on the forms of utterances would be handled by Pickering and Garrod's simple priming account, or by the "output-input coordination" principle of Garrod and Anderson (1987), which predicts that a speaker should continue to use the same expression that worked last time (regardless of addressee). As we argue in the next section, though, there are good reasons to eschew an account of common ground processing that relies on *detailed* metarepresentations of other's knowledge. But, assuming much simpler types of representations for purposes of conversational interaction need not doom individuals to egocentrism. Both of the accounts that we describe next support ways in which "ordinary" partner-relevant representations could give rise to felicitous language use that are consistent with constraints based on cognitive capacities of individual speakers and salient features of conversational contexts.

ORDINARY MEMORY AND "ONE BIT" REPRESENTATIONS

The examples from Stellmann and Brennan's corpus demonstrate that when speakers use a referring expression, they can depend on their addressees to let them know if a referent is unclear. They can take the risk of starting to speak in a timely manner, designing referring expressions based on available representations that are likely (but not guaranteed) to work. If we accept that ordinary conversational reference is unlikely to occur in a resource-intensive manner qualitatively similar to the deliberative consideration of false belief, and if we accept that low-level priming explanations such as output-input coordination (Garrod and Anderson, 1987) or interactive alignment (Pickering and Garrod, 2004) fail to adequately account for audience design, the question remains as to how interlocutors so often are able to refer to objects in ways that are generally consistent with shared knowledge. Here, we consider two accounts that do not see fully elaborated partner models as being necessary for particular referring expressions to succeed—indeed, both emphasize the fact that, in conversation, success is not guaranteed. Specifically, these accounts are the "memory-based" account described by Horton and Gerrig (2005a, in press), and the "one-bit" account described by Galati and Brennan (2006; 2010; see also Brennan and Hanna, 2009).

These two accounts complement each other, in that both connect the dots between memory representations and audience design, with each emphasizing a different launching point: The memory-based account begins with ordinary memory processes and representations, in order to consider how apparent instances of audience design emerge fluidly in conversation, whereas the one-bit account begins with audience design or partner specific processing in conversation, in order to consider what sorts of context-augmented representations (or common ground) could underlie ordinary language processing. Common to both accounts is the idea that partner-specific referring in conversation is mediated through ordinary memory representations (without appeal to special or elaborate mechanisms of the kind entailed by the notion of reference diaries or theory of mind). Both accounts reject as implausible (and computationally expensive) the idea that people routinely "tag" (and continually make triple co-presence inferences about) every element of information that could be relevant to "common ground." Both accounts endorse the view that relevant representations frequently include contextual information concerning a conversational partner, but that there is nothing qualitatively "special" about this information that gives it priority over other relevant information (see Goldinger, 1998 for a similar view).

Moreover, the extent to which person-centered information might be individuated in ways that reliably support felicitous reference will, on both accounts, depend greatly on factors such as immediacy, relevance, perceptual vividness, and goals, as well as whether such information has already been processed in the conversational context. In these respects, both the one-bit view and the memory-based view highlight how routine cognitive considerations strongly shape the extent to which people show evidence for consideration of common ground. On these accounts, successful definite reference depends on simpler representations that support rapid access to contextually relevant knowledge.

The memory-based model (Horton and Gerrig, 2005a, in press) is a cognitively motivated account explaining how language users could gain access to partner-relevant information in ways that require neither special-purpose representations nor special-purpose processes. More specifically, Horton and Gerrig (2005a) described how considerations of common ground could occur on the basis of ordinary representations as they become accessible from memory through ordinary means. For instances of "personal" common ground in particular (Clark, 1996), these memory representations were characterized as rich episodic traces of previous encounters with others, representing the products of the kinds of encoding typical of one's experiences of particular events.

Once such traces are encoded into memory, subsequent encounters can trigger the automatic retrieval of relevant memory traces, through a process termed *resonance*. Inspired by cue-driven retrieval processes found in models of recognition memory (Ratcliff, 1978; Gillund and Shiffrin, 1984; Hintzman, 1986; Ratcliff and McKoon, 1988), resonance involves the parallel activation of information that shares overlapping features with the memory probe (e.g., the presence of an interlocutor). When resonance reaches some threshold, which itself is a function of the recency and frequency with which a memory has been previously retrieved, this information can become accessible in a way that influences other processes. On this account, then, implicit "assessments" of common ground emerge from a speaker's automatic sense that particular information can be treated as familiar or not within a particular context. This in turn can lead speakers to use particular forms of reference if relevant linguistic representations become sufficiently accessible via resonance within a time course that can affect planning and production. Critically, though, under this proposal audience design does not require that these representations be specifically tagged with respect to common ground, although relevant memory traces may still happen to encode partner-relevant information.

The one-bit proposal (Galati and Brennan, 2006, 2010; Brennan and Hanna, 2009) arose from the observation that, in experiments designed to distinguish the perspectives of conversational partners, common ground seems to be able to rapidly guide comprehension and planning of referring expressions when conditions differ by one or just a few well-established, relevant cues that in these experiments happened to be binary—for example, *my partner is a native speaker of English*, or not (Bortfeld and Brennan, 1997); *my partner can see what I'm doing*, or not (Brennan, 2005); *my partner and I have the same spatial perspective*, or not (Schober, 1993; Duran et al., 2011); *my partner can reach the object she's talking about*, or not (Hanna and Tanenhaus, 2004); *my partner can see a picture of what we're discussing*, or not (Lockridge and Brennan, 2002); *I have talked about this with my partner before*, or not (Horton and Gerrig, 2002; Metzing and Brennan, 2003; Galati and Brennan, 2006; Matthews et al., 2010); or *my partner and I were interrupted before we finished discussing this*, or not (Brown-Schmidt, 2009). Such contextual cues, especially if they're established as relevant through perceptual salience or having made a previous inference, can support the rapid access and use of episodic information in order for speakers to design an utterance *for* a particular audience (as opposed to behaving "egocentrically"), or in order for addressees to adapt the processing of a referring expression with a particular speaker in mind.

But even for the simplest of metarepresentations (e.g., *she can/cannot see me talking*), in order for such partner-specific information to guide processing, it must already be accessible (Horton and Gerrig, 2002); this means that the first time an inference is needed, this is likely to require additional processing time. This was demonstrated in a referential communication study (Hwang et al., 2015) in which Koreans who spoke English as a second language worked with a native English speaker to match labels that would be unpronounceable in Korean (which lacks not only any coda-final /b/ vs. /p/ contrast, but also any contrast between the vowels /æ/ and /ɛ/). The Korean speakers did not spontaneously produce recognizable contrasts unless they had just been primed with a similar sound by the native-English-speaking partner, or *unless there was a pragmatic reason to do so*, in order to make a relevant pragmatic distinction that their partners needed to do the matching task—for example, to distinguish a card labeled *bib* from a nearby *bip*. Critically, the first time they encountered their partner's pragmatic need, it took them significantly longer to initiate speaking; but when a similar pragmatic contrast (between different items) was needed after that, they were just as fast to initiate speaking as with other, baseline expressions. This evidence supports the idea that a representation of the discourse context that includes pragmatic information that has already been perceived or computed can rapidly shape referring without the need for elaborate, computationally expensive inferences (see Brennan and Hanna, 2009; Shintel and Keysar, 2009).

Thus, the one-bit account presumes a role for metarepresentations of episodes relevant to conversational contexts, permitting language use to be shaped by inferences that concern what other people might know. In particular, these inferences are most likely to occur either *before* a speaker formulates a referring expression (or an addressee interprets it), based on salient percepts about their physical co-presence, or at the time when a speaker or an addressee is first prompted to consider pragmatically-relevant differences, especially differences that are relatively simple and supported by a stable conversational context. At the same time, however, any metarepresentations that encode such inferences are subject to the constraints of ordinary memory; that is, they are likely to vary in strength and be schematically focused on critical features of the interactive setting. Once evoked, they may support the rapid or automatic use of partner-specific information in similar contexts [as suggested by the Hwang et al. (2015) results described above], rather than requiring additional laborious inference. For example, if an inference about common ground has already been made, or if relevant evidence is perceptually salient, then such information could be available to be used in the formulation and interpretation of referring expressions without further delay (Galati and Brennan, 2006; Brennan and Hanna, 2009). This can result in rapid and "smart" (rather than slow and laborious) adaptations of utterances to a particular partner's needs, perspective, or context.

Critically, the rapid use of partner-specific knowledge under these circumstances could readily occur via the kinds of memory-based processes described by Horton and Gerrig (2005a). Even if metarepresentations are not *always* deployed in language use, the number and variety of findings showing context-appropriate uses of perspective (see discussion in Brennan and Hanna, 2009; Brown-Schmidt and Hanna, 2011) demand that researchers provide a psychologically-plausible account of when and how speakers might come to consider inferences about others' knowledge. For its part, Horton and Gerrig's (2005a) memory-based account does not deny the possibility of metarepresentation. Indeed, in their description of *strategic* assessments of a partner's knowledge, they identified several instances in a corpus of telephone conversations in which speakers appeared to consult (or, more likely, construct) representations of what their addressees might know (e.g., "Yeah, I've got another buddy who, uh, is a Marine pilot. I'm trying to think if you had ever met this guy"). The suggestion, however, was that such activity was likely to be too computationally effortful and/or costly to provide a general account of audience design.

Even so, the primary focus of the memory-based view concerned the automatic accessibility of partner-relevant information via resonance. Much of the time, these partner-relevant representations will be limited to episodic traces of previous encounters with others, allowing individuals to show evidence for partner-sensitivity without requiring detailed inferences about common ground. In principle, though, because resonance is a "dumb" process that works on whatever information is available in memory, there is nothing inherent about resonance as a process that would prevent it from facilitating the retrieval of representations that capture inferences about the knowledge of certain partners—as long as such information is part of the memory trace.

One of the fundamental observations about resonance as an ordinary memory process is that the types of information that become accessible via resonance are likely to be highly dependent upon features of the conversational context, both in terms of encoding strength as well as the availability of appropriate retrieval cues. Thus, certain conversational situations might not only be more likely to result in stronger memory traces for particular sources of information (Horton and Gerrig, 2005b), but might also unfold in a way that supports the direct encoding of highly constrained inferences concerning other's beliefs about that information. In particular, aspects of conversational grounding described by Clark and colleagues could under many circumstances provide the right setting for particular metarepresentational inferences to be encoded as part of the episodic trace for particular interactions (as first suggested by Clark and Marshall, 1981). For example, explicit indications that an individual has understood a particular conceptual perspective (e.g., "alright I see it, I see it" in the opening example) might be more likely to lead to the encoding of the belief that this speaker views that referent in this way. It would be important to empirically distinguish this, though, from the simpler possibility that particular kinds of feedback may just generally lead to stronger memory traces for the interaction.

One piece of evidence on this point comes from Brown-Schmidt (2012), who showed that participants generated stronger inferences about shared knowledge in situations in which they responded to direct questions from a confederate speaker, consistent with the idea that common ground is mediated via "gradient" representations. Thus, while information about simple verbal events such as "Sally called this *an anchor*" would, on any account, almost certainly be present in the episodic trace, processes of negotiating reference might enable information such as "Sally believes this can be conceptualized as an anchor" to (probabilistically) become part of the trace as well. The probabilistic nature of common ground is an underappreciated part of Clark and colleague's original theory, which models the strength of evidence interlocutors provide about their understanding and uptake during conversational interaction (Clark and Schaefer, 1989; Schober and Clark, 1989; Clark and Brennan, 1991; Brennan and Clark, 1996). As meanings are grounded, speakers provide metalinguistic cues as to their commitment to the content of their utterances (Smith and Clark, 1993), and hearers accurately understand and use such cues (Brennan and Williams, 1995; Swerts and Krahmer, 2005). So even though cognitive restrictions prevent individuals from encoding anything that resembles the infinite regress of mutual knowledge, they may be able to use metalinguistic cues and co-presence heuristics to estimate mutual knowledge.

Such cognitive restrictions tie in directly to the motivation for the one-bit account as proposed by Galati and Brennan (2006, 2010). Clearly, ordinary memory representations of particular social interactions cannot encode every possible inference concerning other people and potential referents as part of a simple metarepresentation. But if the conversational context supports a particular inference concerning the likely knowledge of others as relevant, then it is possible that language users may easily encode such inferences as part of the partner-specific memory trace and use it automatically in reference resolution. That is, once so-called metarepresentational information has been computed, it can be available for subsequent retrieval via ordinary resonance, as described by the memory-based model. As such, this retrieval need not be guided in the moment by explicit deliberations about another's perspective.

We wish to stress, though, that the types of representations of another's knowledge likely to be most relevant for everyday language use will most generally be quite different from the sorts of discrete, all-or-nothing inferences about mental states commonly presumed in the literature on theory of mind. That is, we believe that variation across conversational contexts, as well as within conversations over time, will shape the kinds of partner-specific information that become accessible for particular speakers as they formulate and comprehend utterances. Under these circumstances, some aspects of metarepresentational beliefs will be more immediately accessible than others, and any such knowledge is likely to be partial or schematic, depending on the nature of memory cues present in the conversational situation as well as the strengths of the underlying traces stored in memory. With repeated referring, memory traces are stronger, and so any "conceptual pact" achieved by two conversational partners to refer to a particular referent with a particular label is likely to be stronger and more easily evoked (Brennan and Clark, 1996), consistent with the probabilistic or gradient nature of common ground (Brown-Schmidt, 2012). Moreover, what might be seen as computationally expensive (such as keeping track of individual information) need not be so, if ordinary memory processes also support the binding of relevant contextual factors as part of the same representation. But, with highly similar episodes involving context switches (such as interactions with different partners), distinctions between relevant representations might become blurred, leading to further opportunities for interference leading to source-monitoring errors or egocentric mistakes.

Our converging viewpoints emphasize the fact that consideration of another's knowledge is rarely likely to be a discrete, all-or-nothing event, instead unfolding over time as cues become available in the conversational context, leading to the retrieval of partner-focused representations that are incremental and dynamically changing as new information comes online. Furthermore, we suggest that many of these inferences are likely to be simple, reusable, and highly supported by salient features of the conversational context. Once these inferences have been computed and encoded as part of the memory trace for that interaction, the resulting metarepresentations, whether schematized or not, are potentially available for retrieval via ordinary memory-based processes. As a function of ordinary memory, this retrieval will be highly dependent upon the presence of appropriate cues and can still wax and wane as the conversation proceeds.

CONCLUSIONS

By taking seriously questions about the nature of representations to which language users have access for purposes of conversational reference, our aim has been to emphasize the extent to which such representations are constrained in a number of ways by ordinary memory, consistent with constraint-based approaches to reference and common ground (Hanna et al., 2003; Brown-Schmidt and Hanna, 2011). Such constraints are not only internal, tied to fundamental cognitive processes of attention and memory, but also external, arising from the conversational situation (including processes of grounding meanings with a partner). In particular, the issue of metarepresentation highlights key questions about the nature of the ordinary memory traces that potentially encode inferences about another's knowledge. Another person's knowledge or perspective may well have the same status in a metarepresentation as any other relevant aspect of the context in which referring takes place, or it may be supported by distinct neural circuitry (for discussion, see Brennan et al., 2010). Neuroscience has begun to explore the neurocognitive markers associated with particular types of socially relevant capacities, including theory of mind (e.g., Ruby and Decety, 2003; Baron-Cohen et al., 2013; Hamilton et al., 2014). It remains to be seen whether representations of (or inferences about) other people's knowledge, needs, or beliefs are qualitatively different from any other contextual representations or inferences required by language use in conversation.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

AUTHOR NOTES

This material is based on work done while SB was serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Richard Gerrig and the two reviewers for their comments on an earlier draft.

REFERENCES

Apperly, I. (2011). *Mindreaders: The Cognitive Basis of "Theory of Mind."* Hove: Psychology Press.

Apperly, I. (2013). "Can theory of mind grow up: mindreading in adults, and its implications for the development and neuroscience of mindreading," in *Understanding Other Minds: Perspectives from Developmental Social Neuroscience, 3rd Edn.*, eds S. Baron-Cohen, H. Tager-Flusberg, and M. V. Lombardo (Oxford: Oxford University Press), 72–92.

Baillargeon, R., Scott, R., and He, Z. (2010). False-belief understanding in infants. *Trends Cogn. Sci.* 14, 110–118. doi: 10.1016/j.tics.2009.12.006

Baron-Cohen, S., Leslie, A., and Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Tager-Flusberg, H., and Lombardo, M. V. (2013). *Understanding Other Minds: Perspectives from Developmental Social Neuroscience, 3rd Edn*. Oxford: Oxford University Press.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: listeners anticipate but do not integrate common ground. *Cognition* 109, 10–40. doi: 10.1016/j.cognition.2008.07.005

Barr, D. J., and Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *J. Mem. Lang.* 46, 391–418. doi: 10.1006/jmla.2001.2815

Bell, A. (1984). Language style as audience design. *Lang. Soc.* 13, 145–204. doi: 10.1017/S004740450001037X

Bloom, P., and German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77, B25–B31. doi: 10.1016/S0010-0277(00)00096-2

Brennan, S. E. (2005). "How conversation is shaped by visual and spoken evidence," in *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, eds J. Trueswell and M. Tanenhaus (Cambridge, MA: MIT Press), 95–129.

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493. doi: 10.1037/0278-7393.22.6.1482

Brennan, S. E., Galati, A., and Kuhlen, A. K. (2010). Two minds, one dialog: coordinating speaking and understanding. *Psychol. Learn. Motiv.* 53, 301–344. doi: 10.1016/S0079-7421(10)53008-1

Brennan, S. E., and Hanna, J. E. (2009). Partner-specific adaptation in dialogue. *Top. Cogn. Sci.* 1, 274–291. doi: 10.1111/j.1756-8765.2009.01019.x

Brennan, S. E., and Williams, M. (1995). The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *J. Mem. Lang.* 34, 383–398. doi: 10.1006/jmla.1995.1017

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychon. Bull. Rev.* 16, 893–900. doi: 10.3758/PBR.16.5.893

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Lang. Cogn. Process.* 27, 62–89. doi: 10.1080/01690965.2010.543363

Brown, P., and Dell, G. S. (1987). Adapting production to comprehension: the explicit mention of instruments. *Cogn. Psychol.* 19, 441–472. doi: 10.1016/0010-0285(87)90015-6

Brown-Schmidt, S., and Hanna, J. E. (2011). Talking in another person's shoes: incremental perspective-taking in language processing. *Dialogue Discourse* 2, 11–33. doi: 10.5087/dad.2011.102

Bortfeld, H., and Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Process.* 23, 119–147. doi: 10.1080/01638537709544986

Call, J., and Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Dev.* 70, 381–395. doi: 10.1111/1467-8624.00028

Call, J., and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends Cogn. Sci.* 12, 187–192. doi: 10.1016/j.tics.2008.02.010

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., and Brennan, S. E. (1991). "Grounding in communication," in *Perspectives on Socially Shared Cognition,* eds L. B. Resnick, J. Levine, and S. D. Teasley (Washington, DC: APA), 127–149.

Clark, H. H., and Marshall, C. R. (1978). "Reference diaries," in *Theoretical Issues in Natural Language Processing,* Vol. 2, ed D. L. Waltz (New York, NY: Association for Computing Machinery), 57–63.

Clark, H. H., and Marshall, C. R. (1981). "Definite reference and mutual knowledge," in *Elements of Discourse Understanding,* eds A. K. Joshi, B. Webber, and I. Sag (Cambridge: Cambridge University Press), 10–63.

Clark, H. H., and Murphy, G. L. (1982). "Audience design in meaning and reference," in *Language and Comprehension,* eds J.-F. Le Ny and W. Kintsch (Amsterdam: North-Holland), 287–299.

Clark, H. H., and Schaefer, E. F. (1989). Contributing to discourse. *Cogn. Sci.* 13, 259–294. doi: 10.1207/s15516709cog1302_7

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Cohen, P. R., and Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cogn. Sci.* 3, 177–212. doi: 10.1207/s15516709cog0303_1

Dell, G. S., and Brown, P. M. (1991). "Mechanisms for listener-adaptation in language production: limiting the role of the 'model of the listener," in *Bridges between Psychology and Linguistics,* eds D. Napoli and J. Kegl (San Diego, CA: Academic Press), 105–129.

Dennett, D. C. (1978). Beliefs about beliefs. *Behav. Brain Sci.* 4, 568–570. doi: 10.1017/S0140525X00076664

Ditmarsch, H. P. V., van der Hoek, W., and Kooi, B. P. (2007). *Dynamic Epistemic Logic. Synthese Library Series*, *Vol. 337*. Berlin: Springer.

Duran, N. D., Dale, R., and Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive cost. *Cognition* 121*,* 22–40. doi: 10.1016/j.cognition.2011.06.009

Epley, N., Keysar, B., Van Boven, L., and Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *J. Pers. Soc. Psychol.* 87, 327–339. doi: 10.1037/0022-3514.87.3.327

Frith, C. D., and Frith, U. (2012). Mechanisms of social cognition. *Annu. Rev. Psychol.* 63, 287–313. doi: 10.1146/annurev-psych-120710-100449

Galati, A., and Brennan, S. E. (2006). "Given-new attenuation effects in spoken discourse: for the speaker, or for the addressee?," in *Abstracts of the Psychonomic Society, 47th Annual Meeting* (Houston, TX), 15.

Galati, A., and Brennan, S. E. (2010). Attenuating information in spoken communication: for the speaker, or for the addressee? *J. Mem. Lang.* 62, 35–51. doi: 10.1016/j.jml.2009.09.002

Garrod, S., and Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition* 27, 181–218. doi: 10.1016/0010-0277(87)90018-7

Gillund, G., and Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychol. Rev.* 19, 1–65. doi: 10.1037/0033-295X.91.1.1

Goldinger, S. D. (1998). Echoes of echoes? *An episodic theory of lexical access. Psychol. Rev.* 105, 251–279. doi: 10.1037/0033-295X.105.2.251

Gorman, K. S., Gegg-Harrison, W., Marsh, C. R., and Tanenhaus, M. K. (2013). What's learned together stays together: Speakers' choices of referring expressions reflects shared experience. *J. Exp. Psychol.* 39, 843–853. doi: 10.1037/a0029467

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics: Speech Acts,* Vol. 3, eds P. Cole and J. Morgan (New York, NY: Academic Press), 41–58.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: a framework for modeling the local discourse. *Comput. Linguist.* 21, 203–225.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions. *Language* 69, 274–307. doi: 10.2307/416535

Hamilton, A. F., de, C., Kessler, K., and Creem-Regehr, S. H. (2014). Perspective taking: building a neurocognitive framework for integrating the "social" and the "spatial." *Front. Hum. Neurosci.* 8:403. doi: 10.3389/fnhum.2014.00403

Hanna, J. E., Tanenhaus, M. K., and Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *J. Mem. Lang.* 49, 43–61. doi: 10.1016/S0749-596X(03)00022-6

Hanna, J. E., and Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cogn. Sci.* 28, 105–115. doi: 10.1016/j.cogsci.2003.10.002

Hintzman, D. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychol. Rev.* 93, 411–428. doi: 10.1037/0033-295X.93.4.411

Horton, W. S., and Gerrig, R. J. (2002). Speakers' experiences and audience design: knowing *when* and knowing *how* to adjust utterances to addressees. *J. Mem. Lang.* 47, 589–606. doi: 10.1016/S0749-596X(02)00019-0

Horton, W. S., and Gerrig, R. J. (2005a). Conversational common ground and memory processes in language production. *Discourse Process.* 40, 1–35. doi: 10.1207/s15326950dp4001_1

Horton, W. S., and Gerrig, R. J. (2005b). The impact of memory demands on audience design during language production. *Cognition* 96, 127–142. doi: 10.1016/j.cognition.2004.07.001

Horton, W. S., and Gerrig, R. J. (in press). Revisiting the memory-based processing approach to common ground. *Top. Cogn. Sci.*

Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117.

Hwang, J., Brennan, S. E., and Huffman, M. (2015). Phonetic adaptation in non-native spoken dialogue: effects of priming and audience design. *J. Mem. Lang.* 81, 72–90. doi: 10.1016/j.jml.2015.01.001

Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. (2000). Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychol. Sci.* 11, 32–38. doi: 10.1111/1467-9280.00211

Keysar, B., Barr, D. J., and Horton, W. S. (1998). The egocentric basis of language use: insights from a processing approach. *Curr. Dir. Psychol. Sci.* 7, 46–50. doi: 10.1111/1467-8721.ep13175613

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89, 25–41. doi: 10.1016/S0010-0277(03)00064-7

Lockridge, C. B., and Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychon. Bull. Rev.* 9, 550–557. doi: 10.3758/BF03196312

Matthews, D., Lieven, E., and Tomasello, M. (2010). What's in a manner of speaking? *Children's sensitivity to partner-specific referential precedents. Dev. Psychol.* 46*,* 749–760. doi: 10.1037/a0019657

Metzing, C., and Brennan, S. E. (2003). When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *J. Mem. Lang.* 49, 201–213. doi: 10.1016/S0749-596X(03)00028-7

Penn, D. C., and Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything resembling a 'theory of mind.' *Philos. Trans. Roy. Soc. B Biol. Sci.* 362, 731–744. doi: 10.1098/rstb.2006.2023

Perner, J., and Ruffman, T. (2005). Infants' insight into the mind: how deep? *Science* 308, 214–216. doi: 10.1126/science.1111656

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 167–226. doi: 10.1017/S0140525X04000056

Polichak, J. W., and Gerrig, R. J. (1998). Common ground and everyday language use: comments on Horton and Keysar (1996). *Cognition* 66, 183–189.

Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 4, 515–526. doi: 10.1017/S0140525X00076512

Pylyshyn, Z. W. (1978). When is attribution of beliefs justified? *Behav. and Brain Sci.* 4, 592–593. doi: 10.1017/S0140525X00076895

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.* 85, 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., and McKoon, G. (1988). A retrieval theory of priming in memory. *Psychol. Rev.* 95, 385–408. doi: 10.1037/0033-295X.95.3.385

Rubio-Fernández, P., and Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychol. Sci.* 24, 27–33. doi: 10.1177/0956797612447819

Ruby, P., and Decety, J. (2003). What you believe versus what you think they believe: a neuroimaging study of conceptual perspective-taking. *Eur. J. Neurosci.* 17, 2475–2480. doi: 10.1046/j.1460-9568.2003.02673.x

Samson, D., Apperly, I. A., Braithwaite, J., Andrews, B., and Bodley, S. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *J. Exp. Psychol. Hum. Percept. Perform.* 36, 1255–1266. doi: 10.1037/a0018729

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition* 47*,* 1–24. doi: 10.1016/0010-0277(93)90060-9

Schober, M. F., and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cogn. Psychol.* 21, 211–232. doi: 10.1016/0010-0285(89)90008-X

Shintel, H., and Keysar, B. (2009). Less is more: a minimalist account of joint action in communication. *Top. Cogn. Sci.* 1, 260–273. doi: 10.1111/j.1756-8765.2009.01018.x

Smith, N. V. (1982). *Mutual Knowledge*. London: Academic Press.

Smith, V. L., and Clark, H. H. (1993). On the course of answering questions. *J. Mem. Lang.* 32, 25–38. doi: 10.1006/jmla.1993.1002

Sperber, D. (2000). *Metarepresentations: A Multidisciplinary Perspective*. Oxford: Oxford University Press.

Stellmann, P., and Brennan, S. E. (1993). "Flexible perspective-setting in conversation," in *Abstracts of the Psychonomic Society, 34th Annual Meeting* (Washington, DC), 20.

Surtees, A., and Apperly, I. A. (2012). Egocentrism and automatic perspective-taking in children and adults. *Child Dev.* 83, 452–460. doi: 10.1111/j.1467-8624.2011.01730.x

Swerts, M., and Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *J. Mem. Lang.* 53, 81–94. doi: 10.1016/j.jml.2005.02.003

van der Wel, R. P. R. D., Sebanz, N., and Knoblich, G. (2014). Do people automatically track others' beliefs? *Evidence from a continuous measure. Cognition* 130, 128–133. doi: 10.1016/j.cognition.2013.10.004

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory of mind development: the truth about false belief. *Child Dev.* 72, 655–684. doi: 10.1111/1467-8624.00304

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128. doi: 10.1016/0010-0277(83)90004-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 April 2016
doi: 10.3389/fpsyg.2016.00561

The Flexibility of Conceptual Pacts: Referring Expressions Dynamically Shift to Accommodate New Conceptualizations

Alyssa Ibarra[1,2]* *and Michael K. Tanenhaus*[1,2]

[1] Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA, [2] Department of Linguistics, School of Arts & Sciences, University of Rochester, Rochester, NY, USA

In a classic paper, Brennan and Clark argued that when interlocutors agree on a name for an object, they are forming a temporary agreement on how to conceptualize that object; that is, they are forming a conceptual pact. The literature on conceptual pacts has largely focused on the costs and benefits of breaking and maintaining lexical precedents, and the degree to which they might be partner-specific. The research presented here focuses on a question about conceptual pacts that has been largely neglected in the literature: To what extent are conceptual pacts specific to the local context of the interaction? If conceptual pacts are indeed temporary, then when the local context changes in ways that are accessible to participants, we would expect participants to seamlessly shift to referential expressions that reflect novel conceptualizations. Two experiments examined how referential forms change across context in collaborative, task-oriented dialog between naïve participants. In Experiment 1, names for parts of an unknown object were established in an "item" identification stage (e.g., a shape that looked like a wrench was called "the wrench"). In a second "build" stage, that name was often supplanted by an object-oriented name, e.g., the "leg." These changes happened abruptly and without negotiation. In Experiment 2, interlocutors manipulated clip art and more abstract tangram pictures in a "slider" puzzle to arrange the objects into a target configuration. On some trials moving an object revealed a picture that could be construed as a contrast competitor, e.g., a clip art picture of a camel after "the camel" had been negotiated as a name for a tangram shape, or vice versa. As would be expected, modification rates increased when a potential contrast was revealed. More strikingly, the degree to which a name had been negotiated or the frequency with which it had been used did not affect the likelihood that the revealed shape would be considered as a potential contrast. We find little evidence that names that are introduced as part of a conceptual pact persist when either the task goals or informational needs change. Rather, conceptual pacts are fluid temporary agreements.

Keywords: referential expressions, conceptual pacts, entrenchment, targeted language game, interactive conversation

OPEN ACCESS

Edited by:
Albert Gatt,
University of Malta, Malta

Reviewed by:
Laura Gonnerman,
McGill University, Canada
Paula Rubio-Fernández,
University of Oslo, Norway

****Correspondence:***
Alyssa Ibarra
aibarra@ur.rochester.edu

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

Received: *10 September 2015*
Accepted: *05 April 2016*
Published: *25 April 2016*

Citation:
Ibarra A and Tanenhaus MK (2016)
The Flexibility of Conceptual Pacts:
Referring Expressions Dynamically
Shift to Accommodate New
Conceptualizations.
Front. Psychol. 7:561.
doi: 10.3389/fpsyg.2016.00561

INTRODUCTION

There is a many-to-many mapping between names and potential referents. A picture of a Bernese Mountain Dog could be referred to as “the dog”, “the Bernese Mountain Dog,” “the Berner,” etc. Furthermore, names can be assigned on the basis of the properties of the object (e.g., “the brown dog”), assigned with respect to a particular referential domain (e.g., “the big dog” when there is a small one as well), or assigned with respect to a referential domain that is tied to a local goal. For example, when completing a jigsaw puzzle, “try the red piece,” might be uttered—and easily understood—in a collaborative task, when there are several pieces that might fit and several other red pieces that are clearly the wrong size and shape (Brown-Schmidt and Tanenhaus, 2008). A central question, then, in both reference generation and comprehension is how interlocutors choose a specific referential expression and then modify its use as a discourse unfolds.

In a classic paper, Brennan and Clark (1996) introduced the notion of a *conceptual pact*. Building on the foundation established by Clark and Wilkes-Gibbs (1986) and Isaacs and Clark (1987) in developing the collaborative model of conversation, Brennan and Clark argued against what they termed “a historical” accounts of the generation of referring expressions. They proposed that when participants agree upon a name, that is when the name has been *grounded*, they are making a *temporary* agreement about how to conceptualize that object. The aspect of conceptual pacts that has received the most attention in recent years is the claim that conceptual pacts are partner-specific. For example, there is an ongoing debate about how and when the use of a particular name for an object will affect processing, specifically when an established name (a maintained precedent) or a different name (a broken precedent) is used by the same partner or by a different partner (see Kronmüller and Barr, 2015 for a recent review and meta-analysis).

The current research focuses on an important aspect of conceptual pacts that has been largely ignored in the literature. In comparing their proposal with pioneering work by Carroll (1980, also see Carroll, 1985), Brennan and Clark noted that a defining characteristic of their model is that conceptual pacts are temporary and adaptable. They argued that participants can modify their utterances abruptly, that is without negotiation in response to changing goals and informational needs. Thus when the interlocutors’ goals change or when the informational demands of the local context change, then the pressure to adhere to a conceptual pact might be relatively weak. If that is the case, then elucidating these factors will be crucial to developing robust theories of reference generation and understanding. Here we ask (1) whether interlocutors do indeed abruptly change their referential expressions–in particular negotiated names–when either the goals or informational demands change; and (2) whether frequent use of a name in a conceptual pact causes the name to become entrenched, that is resistant to being changed.

Given the focus in the literature on the benefits of maintaining precedents and the costs of breaking precedents, Brennan and Clark’s (1996) evidence for conceptual pacts becoming entrenched is arguably quite weak. They found that when interlocutors used a subordinate-level name because a picture was introduced in the context of another member of the same category (e.g., a collie in the context of two dogs), they often continued to use that name rather than the more basic-level name (e.g., “dog”). This can be viewed as a form of over-modification, which is quite common (see Pogue et al., 2016).

Conversely, the evidence for the flexibility of conceptual pacts is also relatively weak. Brennan and Clark (1996) found that after using a basic-level name, such as “shoe,” interlocutors would switch to a subordinate level category, e.g., “penny loafer” when confronted by a situation in which there was more than one type of shoe. In this situation each of the potential referents would be an equally good fit to the previous referential description of “shoe.” Moreover, changing the name from a basic-level category name to a subordinate-level name does not require a major shift in conceptualization.

If a grounded name reflects a *temporary* agreement to conceptualize an object in a particular way with respect to a *particular set of goals and informational needs,* then when the goals or informational needs change, participants will no longer seek to use previously grounded referring expressions. The current research was designed to test this claim using stronger manipulations than those used by Brennan and Clark (1996). In Experiment 1, we introduce a change in the goal structure of a collaborative task that could potentially trigger a bigger shift in conceptualization in order to see whether or speakers will make non-negotiated, abrupt changes in their use of a referring expression. In Experiment 2 we introduce a change in the referential context by introducing a new object that could be treated as contrastive or not, depending on whether participants choose to maintain a previous conceptual pact.

Both experiments use “targeted language games” (Brown-Schmidt and Tanenhaus, 2008; Tanenhaus and Brown-Schmidt, 2008) with task-oriented dialog. In most experimental work on naming, the referential domain depends primarily on a fixed set of potential referents, with little or no collaboration and minimal task constraints (e.g., Nadig and Sedivy, 2002; Sedivy, 2003; Brown-Schmidt and Tanenhaus, 2006; Engelhardt et al., 2006; Brown-Schmidt and Konopka, 2008). Thus there are no changes in either the goals or informational demands that might lead participants to abandon previously established conceptual pacts.

In task-oriented dialog, however, interlocutors work together to establish a given name in order to uniquely identify a referent with respect to goals, which might be hierarchically arranged. Therefore the language is unscripted and the referential domains are likely to be more fluid than in studies using other tasks.

An example of how task constraints and local goals can modulate referential domains comes from a study by Brown-Schmidt and Tanenhaus (2008). In their study, two participants who could not see one another collaborated to match place pieces on their respective Duplo^{TM} pegged boards such that the boards would match. Participants had identical boards that were divided into several sub-regions defined by cardboard borders. Additionally, participants had stickers covering the pegs on the board indicating the type of block (e.g., color and shape) that was to be placed in that particular location; where one participant had a sticker, the other participant's board was blank. Brown-Schmidt and Tanenhaus analyzed utterances in which one of the participants referred to a block when there was a potential competitor referent in the same subregion (e.g., a red, vertical block, when there was another red block, like a red horizontal block). Whereas one might expect speakers to use referential expressions that would take into account all of the potential referents, more than 50% of there referential descriptions were "under-informative," For example, the speaker might say "Put it above the red block." Surprisingly, these potentially ambiguous referring expressions were not confusing to the addressee. When the task constraints were factored in, these "under-informative" referring expressions did in fact uniquely identify a single block, (e.g., only one of the red blocks had enough space above it to put another object). The goal of placing blocks in particular configurations thus restricted the referential domain to only those that fit.

We explore two hypotheses about the context-dependence of conceptual pacts. The first hypothesis is that a conceptual pact is specific to a task goal. The strongest form of this hypothesis is that when the goal changes, interlocutors will neither seek to maintain, nor avoid breaking, lexical precedents. The second hypothesis is that participants will no longer be bound by prior conceptual pacts whenever there is a change in the potential informational demands. We assess this by introducing new objects that could be named in different ways to see if participants seek to generate a referential description that would allow them to maintain a lexical precedent.

EXPERIMENT 1: CONCEPTUAL PACTS AMID SHIFTING GOALS

In Experiment 1 we created a situation in which two interlocutors who could not see each other's work areas collaborated to build a three-dimensional puzzle of an animal. The partners' task was divided into two phases. In the first phase, participants retrieved and named relevant puzzle pieces (abstract shapes) from a larger set using instructional cards as a guide. Each card showed a picture of an object that would be needed later. Participants had no knowledge of the end goal of the puzzle. In the second phase, participants used those pieces to build the animal puzzle. We focused on whether the shift in goal from picking out abstract objects from a set to the assembly of an animal would predict a change in references to objects with previously grounded referring expressions. We asked whether an abstract object referred to by a grounded name (e.g., the wrench) would be referred to by a different, object-based name (e.g., the leg) when the goal is shifted, and if so, whether the change in name would need to be negotiated or even commented on by the interlocutors (e.g., "the wrench is a leg"). We also hoped to find cases where that piece did not fit (e.g., the piece referred to as "the leg" wouldn't fit) in order to see whether participants would then reuse the negotiated name.

Materials and Methods

We created a turn-taking language task designed to elicit repeated reference to a specific set of objects. The design is a subset of a larger paradigm used in studies assessing linguistic convergence and miscommunication (Roche et al., 2013; Paxton et al., 2014). The participants' overall task was to build three-dimensional puzzles using pictorial instruction cards. The task constrained both the objects requiring linguistic reference and the order in which they were to be manipulated, but participants were otherwise able to speak freely. In addition, a barrier was introduced so that there was no shared visual space, thereby necessitating that using spoken language was the primary means of communication (e.g., shifts in gaze or point could not be used as cues). Turns were implemented rigidly, as instructions were staggered across the two instruction stacks, one for each participant and both of which made up the full set of instructions. These were ordered to ensure successful alternation of turns.

Stimuli included three Bloco^{TM} animal puzzles, consisting of a grasshopper, lion and lizard (see **Figure 1** for final product) and order of objects was counterbalanced across participants. The number of instructions varied between the objects (total range: 17–32 instruction cards; see **Table 1** for a breakdown; see **Figure 2** for sample instruction cards).

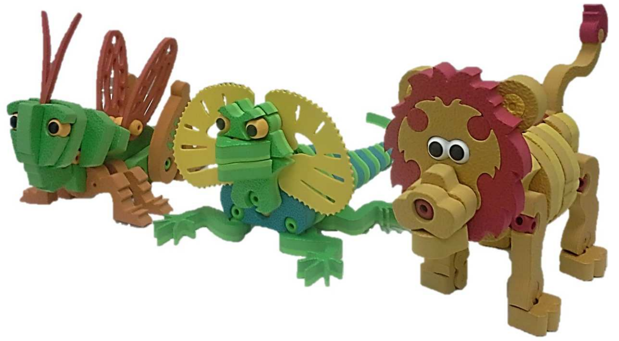


FIGURE 1 | Bloco^{TM} objects: grasshopper, lizard, lion.

TABLE 1 | Number of instruction cards for each animal, total and by phase.

Object	Total instructions	Item phase	Build phase
Lizard	17	9	8
Grasshopper	23	10	13
Lion	32	15	17

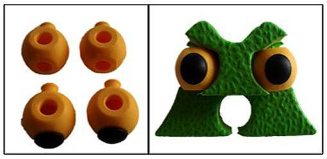


FIGURE 2 | Sample item phase instruction card (left) and build phase instruction card (right) for lizard.

Task: Phase Structure

To create a situation in which repeated reference to the same objects would occur across two different contexts by virtue of a shift in task, separate phases were introduced: the *item phase* and the *build phase*. In the item phase, participants picked out the individual pieces needed to build the object. Both participants were given identical sets of pieces in a bin. During a turn, each participant would pick out the piece on her card from her bin and then describe the object so that her partner could do the same. Instructions alternated until the proper subset of items had been selected. Note that the identity of the object to be built was not given until the item phase was complete.

Once the item phase was complete and the participants had matching pieces on their workspaces, the build phase began. During the build phase, participants also alternated instructions, with each instruction card describing how to build a combination of those pieces obtained in the item phase; these were ordered and staggered to establish turn taking. Crucially, these pieces were to be combined and then added onto a pre-constructed object—an unfinished animal body. The unfinished animal body was presented at the beginning of the build phase in order to emphasize the shift in task. In the item phase, we expected participants were likely to uniquely identify pieces relative to the set of alternatives in the bin. In the build phase, we expected that participants would refer to pieces relative to their relationship to the build-object, especially once its structure began to emerge.

For both phases, participants were video-recorded and the audio was extracted to fully transcribe the participants' dialog. Transcriptions were then coded with the following annotations.

Annotations

Discourse features:

(1) Speaker identity: the identity of the speaker at any given turn.
(2) Local shifts in task: tracking of trial; each instruction card representing a trial, constrained by task and specific object.
(3) Global shifts in task: tracking of phase; either item or build.

Features of referential expression:

(1) Referent: the object being referred to by the referential expression.
(2) Groundedness: presence of grounding expression following referential expression.
Fine-grained: classed individually as "mhm," "yup," "yeah," "yes," "okay," "alright."
Course-grained: classed as "mhm," "yup," "yeah," "yes," "okay," "alright."
(3) Definiteness of description: presence of a definite article preceding description.
(4) Distance from previous mention: number of turns between repeated references to same object, semi-constrained by task.
(5) Noun phrase head: head noun of referential expression.
(6) "Like" complement: referential expression following a "like" construction, most commonly a "looks like" construction.

Outcome Variables

The variable of interest was a change of a referential form, as realized in the repeated reference of any given puzzle piece. This was operationalized as a change in the head noun of a referring expression.

Secondarily, the specific type of change was explored, with any given change categorized as either a *negotiated change* or an *abrupt change*. We operationalized negotiated changes as those referential expressions whose changed form was introduced obliquely in prior discourse but not as the head noun. Because these often were the complements of constructions such as "looks like a wrench," this was the only linguistic construction used to code negotiated changes. This was a simplification, as changes in form might be negotiated in discourse in a different way, however, for present purposes, we restricted negotiated instances to the "like-complement" cases. Abrupt changes were taken to be any change in head noun whose new form was not given in previous discourse. See **Figure 3** for an excerpted transcript of participants who call a piece "the wrench" repeatedly in the item phase but shift to calling it "the leg" without overt negotiation.

Participants

Nine pairs of undergraduates participated from the University of Rochester ($N = 18$). All participants were native speakers of American English with normal to corrected vision. None reported speech or hearing impairments. This study was carried out in accordance with the Research Subjects Review Board (RSRB) at the University of Rochester. All participants gave written informed consent in accordance with the Declaration of Helsinki and were debriefed and offered a copy of the consent form upon completion of the experiment.

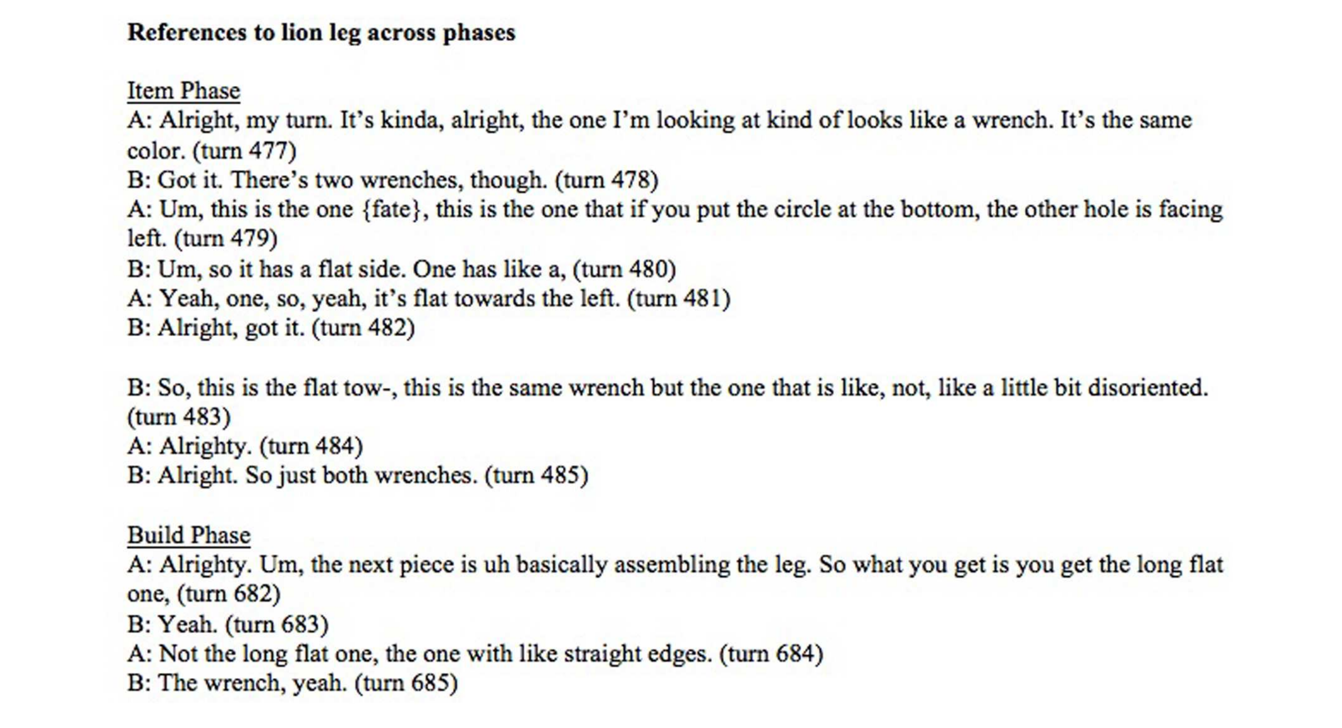
Results

Descriptive Data

The following description and analysis includes data from the nine pairs of participants. The dataset consisted of a total of 1,776 turns containing referential expressions. These referential expressions were hand-annotated and then coded for *groundedness*. A referential expression was coded as grounded if it was followed by any of the following agreement markers: *mhm*, *yup*, *yeah*, *yes*, *okay,* or *alright*. All other referential expressions were coded as not grounded. A non-grounded expression, if taken to be a proposal for a name that has not been accepted, is likely to undergo changes until a name is established. This was not a change in form that we were interested in, since they would not be broken pacts. Therefore, we restricted our dataset to grounded expressions. These made up the subset of data that was used for the following analyses. **Table 2** presents a breakdown of the number of grounded and changed referential expressions. Note that when looking at head noun changes generally, they were quite evenly distributed across groundedness: 50.58% of head noun changes were grounded. Thus changes were not primarily the result of a pact not having been formed in the first place.

Analyses

To assess whether a context shift predicted a change of form, we used a generalized linear mixed model including random



References to lion leg across phases

Item Phase

A: Alright, my turn. It's kinda, alright, the one I'm looking at kind of looks like a wrench. It's the same color. (turn 477)
B: Got it. There's two wrenches, though. (turn 478)
A: Um, this is the one {fate}, this is the one that if you put the circle at the bottom, the other hole is facing left. (turn 479)
B: Um, so it has a flat side. One has like a, (turn 480)
A: Yeah, one, so, yeah, it's flat towards the left. (turn 481)
B: Alright, got it. (turn 482)

B: So, this is the flat tow-, this is the same wrench but the one that is like, not, like a little bit disoriented. (turn 483)
A: Alrighty. (turn 484)
B: Alright. So just both wrenches. (turn 485)

Build Phase

A: Alrighty. Um, the next piece is uh basically assembling the leg. So what you get is you get the long flat one, (turn 682)
B: Yeah. (turn 683)
A: Not the long flat one, the one with like straight edges. (turn 684)
B: The wrench, yeah. (turn 685)

FIGURE 3 | Transcript of participants referring to a puzzle piece that will become the lion leg in the build phase (Pair 4).

TABLE 2 | Total number of referring expressions.

Total referential expressions	Grounded	Changed
871	429	173

Note: A subset of the total were grounded, and a subset of the grounded expressions underwent a change in form (nested values).

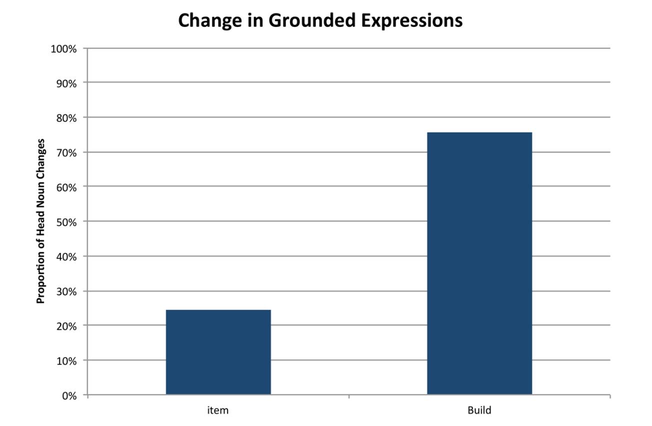


FIGURE 4 | Proportion of changes in grounded referential expressions across phases.

intercepts only, with presence of head noun change as the outcome variable, phase as predictor and pair as a random effect. The model showed that build phase significantly predicts a change in form ($\beta = 0.86$, $SE = 0.13$, $p < 0.001$). **Figure 4** presents the proportion of changes in the two phases.

A second generalized linear model was used to assess how changes were realized across a task-goal shift. We included random intercepts and the model contained phase as the outcome variable, change type (either negotiated or abrupt) as the predictor and pair was the random effect. The decision to reverse the directionality of the model from the first was motivated by the fact that the categories of change type were *ad hoc* categories meant to provide a finer-grained description of a change in form. Thus, we wanted to assess whether occurrences of a particular type of change could predict whether the pair was in the item or build phase. Indeed, the model shows that an occurrence of an abrupt change increases the likelihood that a pair is in the build phase ($\beta = 0.975$, $SE = 0.23$, $p < 0.001$). See **Figure 5** for the proportions of abrupt changes for each phase. Although there were not sufficient trials for a statistical analysis, we observed a striking phenomenon that further highlights the context-dependence of names. On 16 trials, one of which is illustrated in **Figure 3**, participants made an abrupt change from a negotiated name (e.g., *wrench*) to an object-oriented name (*legs*) and then found the piece would not fit. On each of these trials, participants then reverted to the name used during the item-phase (e.g., wrench). This illustrates that participants can shift seamlessly between different names, which reflect different conceptualizations of the objects tied to different task goals.

Discussion

In Experiment 1, changes in the context were instantiated by a shift in task goals, i.e., the shift from the item phase to the

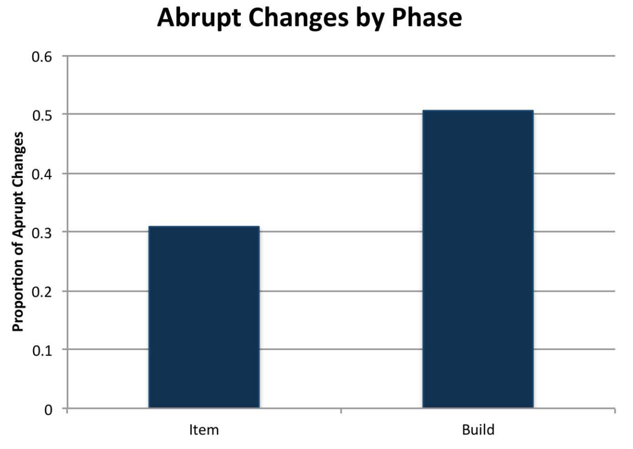


FIGURE 5 | Proportion of abrupt changes in grounded referring expressions within each phase. Negotiated changes are the converse.

build phase. This allowed us to determine whether a conceptual pact that was established when the participants had a particular goal would need to be renegotiated when the goal changed. The answer was clearly "No," highlighting that conceptual pacts are strongly dependent on task goal. We note that an overall task goal, such as building an animal, is likely to have hierarchically organized subgoals, e.g., building a part of the animal (e.g., the face). We are assuming the conversation takes place against the backdrop of these goals as well as more local subgoals. This raises the important issue of how local task goals interface with a conversation as it unfolds. One possibility is that local goals might play a role in determining the Question Under Discussion (Roberts, 1996), but this is beyond the scope of the current paper.

In Experiment 2, we examine the extent to which a name agreed upon in a conceptual pact is entrenched when new objects are introduced that may potentially change the informational demands. First, we explore whether negotiation of a name is predicted from the properties of an object, i.e., whether it is a commonly identifiable object given the alternatives. This manipulation was important for two reasons. First, it allowed us to establish that the typical pattern for negotiated referring expressions, namely a reduction in length across repeated references would be replicated in our task. Secondly, increased negotiation for less identifiable objects would provide evidence that that object is likely to be more compatible with multiple descriptions compared to the more identifiable objects. As will become clear this property of our design creates situations in which the speaker could generate a referential expression that would allow them to maintain a lexical precedent.

EXPERIMENT 2: CONCEPTUAL PACTS AMID EMERGENT NAMING COMPETITORS

We created a targeted language game (Brown-Schmidt and Tanenhaus, 2008; Tanenhaus and Brown-Schmidt, 2008) modeled on puzzles in which a player moves tiles in a constrained space to match a target pattern. Participants were tasked to move the tiles collaboratively to achieve the target configuration. Crucially, three of the tiles were occluded and were only revealed after they had been moved to a particular location. In the critical case, an object that could be identified with a name or a description (e.g., a tangram animal) is occluded and later revealed, allowing for the establishment of a referring expression for the already-visible and more prototypical picture (e.g., a clip art animal). This is contrasted with the reverse order, in which a prototypical picture is visible and a tangram is revealed. Adherence to a conceptual pact was assessed by measuring rate of modification. Unlike in Experiment 1, this set-up calls for the goal and task structure to remain the same and instead assesses the effects of potential changes in information demands. In particular we ask: (1) whether a referential expression for an old object will be modified after encountering a new object and (2) whether the frequency of use, or extent to which a name has been negotiated, affects whether participants continue to use the old name.

Methods and Materials

The game board consisted of a fixed three by three grid. Each cell contained a game tile except for the middle cell, which was empty. **Figure 6** presents a sample game board. Tiles can only be moved into the empty square. Movements into the empty space were possible only for adjacent tiles (e.g., it was invalid to move from top left corner to bottom right corner). The target configuration required the matching of colors and patterns as described below.

Animal Cards

Each game tile in the occupied cells contained a depiction of an individual animal in its center. Half of the cards (i.e., four of the eight) had depictions that were tangrams, or geometric figures, that resemble an animal. The other four cards had more-identifiable clip art animal depictions. To create an image bank of animal graphics of both types, tangram images were used as a base and modified in an image editor by "photo-shopping" in clip-art details to create the more identifiable versions, maintaining color and size consistency in the process. The full image bank consists of 14 animals each with two versions: tangram and clip-art/modified. These are hereafter referred to as *potential naming competitors*. **Figure 6** shows this mix of item types in a sample game board and **Figure 7** illustrates side-by-side comparison of image type. This feature of the game emerged from a starting observation: it is unclear to what degree the entrenchment of a name depends on properties of an object. That is, if partners collaboratively form a conceptual pact to refer to an object by a specific name, the process by which that name is agreed upon might be different depending on the identifiability of the object. By adding tangram versions of items in this game, we created a condition where those objects were not only likely to require more negotiation to establish a name, but could be referred to in different ways due to not being good exemplars of a particular animal.

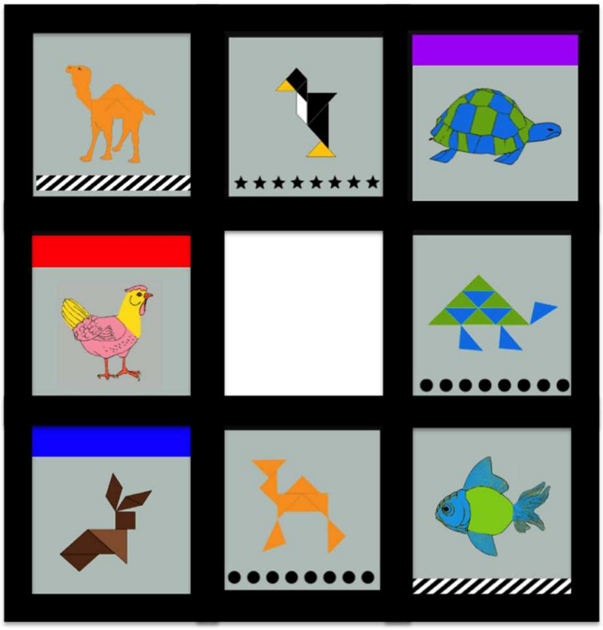


FIGURE 6 | Game board with four tangram images and four clip art images.

Two-Player Adaptation

This game was adapted for two people in the following ways. First, players only have visual access to their own individual game board. They sit face to face with their partners, but each has a computer screen the other cannot see. Secondly, players must play together but make move in alternating order, with one player making a move and instructing her partner to do the same and vice versa. Crucially, each player's game board only has partial information of what is needed to achieve the target configuration: both players' game tiles have the same animal depiction, but there is also accompanying card information that is either displayed in the form of a *color bar* or a *pattern bar*, above or below the central animal. There are three possibilities: Color bars could be red, blue or purple; pattern bars could be stripes, dots, or stars. Furthermore, these bars appear in complementary relation. For example, if Player A's chicken has a red bar on top, Player B's chicken might have stars. Any other animal with a red bar for that game would have a counterpart with stars. Players do not know which color corresponds to which pattern, so they are likely to refer to a game tile by the shared information, i.e., the animal name. Furthermore, displays for both players were mirror images to allow for the same moves while discouraging location-based referring expressions.

In sum, then, the participants' task was to configure the three corresponding red/star animals in one row, the three blue/dots in another and the two purple/stripes in the last row, with the exact configuration left up to the players.

Occluded Tiles

As described above, eight tiles had animal images, but only a subset was visible from the start of the game. In particular, three of the eight tiles initially were hidden by an opaque colored square that covered the animal and color/pattern information. These occluded tiles were revealed when they were moved to the (empty) middle square. Once the occluded tile was moved to the center, the game tile was revealed and remained visible throughout the rest of the game. Of most importance, for the critical game, two of the occluded game tiles covered animal images that were potential naming competitors to animals on the game board: one was a hidden tangram version, the other a hidden clip-art version. The third occluded tile covered a singleton animal to discourage participants from becoming aware that there might be a tangram/clip-art contingency. This allowed for participants to establish names for the visible objects before encountering an alternative that might plausibly be treated as a contrast. More specifically, it created a situation in which a tangram object that after negotiation had been called "the camel" might be moved, and thus referred to, several times before a more prototypical camel is revealed. This set up allowed us to address

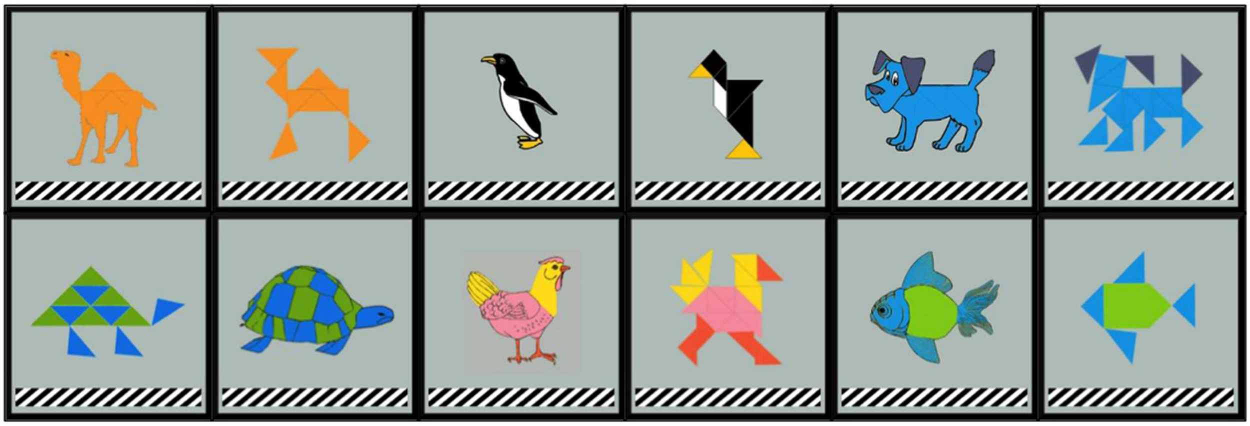


FIGURE 7 | Side by side comparison of tangram and clip art animals. A subset of six is shown.

three questions within the context of an unscripted interaction among naïve participants.

The first question is to what extent a name remains entrenched after a new referent is introduced that could be conceptualized as a contrast to a previously named referent. If the introduction of a naming competitor results in a modification to an established name, this suggests a new conceptualization of the original object, and more importantly, the flexibility of a conceptual pact to accommodate new contextual information. The second question is whether the type of competitor would predict the likelihood of modification. That is, if the revealed object were a tangram, would its name likely be negotiated to a common name without appeal to a contrast set? Similarly, would a hidden clip-art image elicit more modification given that its name is more commonly shared? The third question is whether the name for an object becomes more entrenched with repeated use, i.e., whether it becomes more resistant to modification.

Conditions

Each pair played a total of two games. The first game had all singleton items with no naming competitors occluded or visible. The second game had the crucial occluded tiles covering the tangram and clip art counterparts to visual tiles, as well as a singleton.

To manipulate the relative strength of a conceptual pact we presented the three conditions to also assess carryover across each game, as follows. Game 1 contained all singleton items. Game 2, on the other hand, had counterpart items as well as potential carry over from Game 1. The reason for only including singleton items in Game 1 was to prevent participants from beginning Game 2 with the assumption that occluded tiles might contain potential contrast items.

In the *No Carryover* condition: Game 1 had no items that carried over to Game 2.

In the *Carryover Strong* condition: Game 1 had two singleton items that carried over to Game 2. In Game 2, these singletons were visible and their naming competitors, never before seen, are occluded. This was to encourage a strong conceptual pact for the name given in Game 1.

In the *Carryover Mediated* condition: Game 1 was identical to Game 1 of *Carryover Strong* condition. The same two singleton items carried over to Game 2, however, the items previously seen were the naming competitors that were occluded. They were only revealed later in the game. This is designed to test for reuse of common names for different referents at the start of Game 2.

Annotations for all token referential expressions were isolated and coded for length of expression in number of words, modification (presence/absence), negotiation (presence/absence), naming competitor status (had competitor/did not), ordering (hidden/visible). One pair's Game 1 was not included in the analysis due to a technical error: recording of the Game 2 dialog erased the recording of the Game 1 dialog.

Participants

Fifteen pairs of participants were recruited from the University of Rochester ($n = 30$). All participants were native speakers of American or British English with normal to corrected vision. None reported speech or hearing impairments. One participant reported color-blindness. This study was carried out in accordance with the RSRB at the University of Rochester. All participants gave written informed consent in accordance with the Declaration of Helsinki and were debriefed and offered a copy of the consent form upon completion of the experiment. Speech was recorded and transcribed as a full corpus similar to Experiment 1.

Results

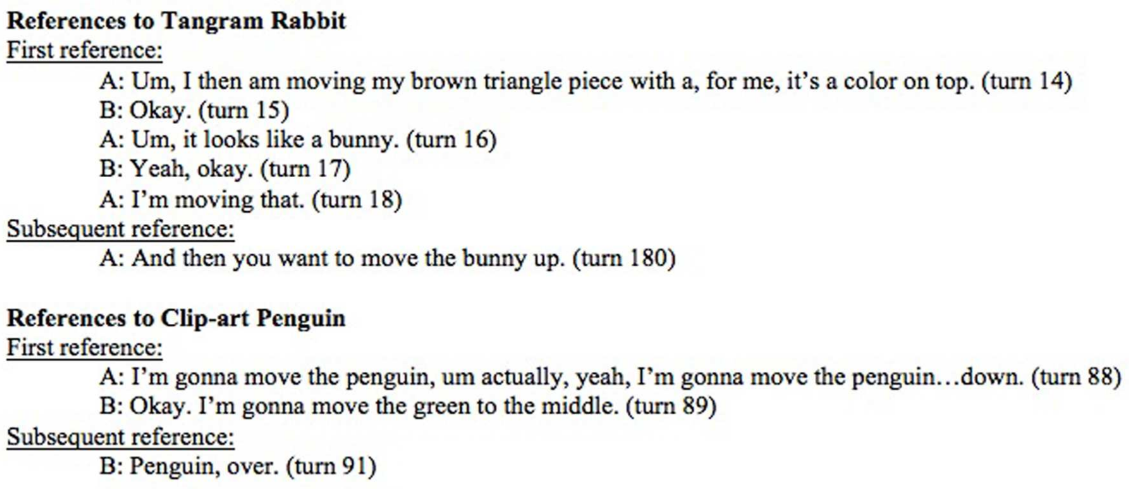
For purposes of the analyses reported here, we collapsed across the three "carryover" conditions. In follow-up work with more pairs, we will examine the effects of these conditions. There were 1671 total referential expressions referring to the animals on the game cards. References to the occluded squares and location-based references to the game cards were omitted from the count. Both occluded squares and location-based references were often referred to with modification (e.g., "the green occluded," "the bottom middle," respectively), which would artificially increase the modification rates of interest. The latter were omitted also because these references had different referents for each partner. Because game boards were displayed as mirror images, participants who used location-based references were often likely to miscommunicate and would only realize later in the game that their partner was not moving the same card she moved. These occurred rarely, except for one pair who did not understand the goal of the game and used these references extensively in Game 1.

Negotiation

Tangram objects were more likely to have names that required negotiation than others. Of the 113 referential expressions that were negotiated, 91.2% were tangrams. See **Figure 8** for a sample of a transcript illustrating one instance of negotiation for a tangram animal but no negotiation for a clip-art animal. In order to measure the strength of association between item type (i.e., whether a card was a tangram or a clip-art graphic) and negotiation status (i.e., a binary yes or no), we used a logistic regression model with negotiation as the outcome variable and item type as a predictor. Results indicate that tangrams have 11.31: 1 odds of negotiation relative to clip-art items (Wald-statistic = 7.24, $p < 0.0001$), with the overall model significant ($p < 0.0001$).

Length of Referring Expression

To assess changes in referring expression, we first measured length of referring expression. We asked whether the utterances in our tasks exhibit the standard behavior of repeated reference in past studies (e.g., Carroll, 1980; Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996). As interlocutors converge on a name, subsequent references should become shorter. **Figure 9** shows the average length of the first three references to an object with a competitor *prior* to the reveal of its competitor. A significant difference of length is shown between clip art and tangram items on the first reference but by the third reference, the average length of referring expression for both item types has converged to one word (e.g., a bare noun). We fit a generalized linear model with



References to Tangram Rabbit

First reference:

A: Um, I then am moving my brown triangle piece with a, for me, it's a color on top. (turn 14)
B: Okay. (turn 15)
A: Um, it looks like a bunny. (turn 16)
B: Yeah, okay. (turn 17)
A: I'm moving that. (turn 18)

Subsequent reference:

A: And then you want to move the bunny up. (turn 180)

References to Clip-art Penguin

First reference:

A: I'm gonna move the penguin, um actually, yeah, I'm gonna move the penguin…down. (turn 88)
B: Okay. I'm gonna move the green to the middle. (turn 89)

Subsequent reference:

B: Penguin, over. (turn 91)

FIGURE 8 | Transcript of participants playing tangram/clip-art game (Pair 6). Negotiated vs. non-negotiated name use.

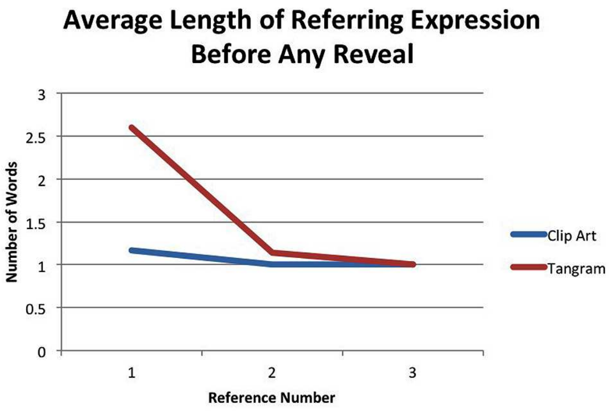


FIGURE 9 | Average number of words for the first three references prior to the reveal of a new referent.

length as the outcome and item type and reference number as the predictor. There was a main effect of both item type and reference number: tangrams were found more likely to have longer lengths overall (β = 0.723, *SE* = 0.2843, $p < 0.05$), and crucially, as references increased, length decreased (β = −0.4539, *SE* = 0.1846, $p < 0.05$).

Next, we assessed length of referring expression before and after the reveal. After the reveal of a competitor item, one might see a similar reduction of the length of the referring expression, indicating further convergence on a name. However, if a participant conceptualizes the revealed object as a contrast item, the length of the referring expression might plateau or even increase. We did a contingency analysis for items with a potential contrast and fit a generalized linear model with length as the outcome and item type and before-and-after reveal status as predictors. There was no main effect of item type ($p = 0.13$) but a main effect for reveal status. That is, for both clip art and tangram objects there was an increase in length after the reveal of a potential contrast (β = 0.7065, *SE* = 0.2462, $p < 0.01$). **Figure 10** shows the average length of referring expression for the single reference before and after the reveal of its competitor.

Length around the reveal of a non-specific item (i.e., an item that is not a competitor) was also assessed. **Figure 11** shows the average length of referring expressions for items that would eventually have a competitor, before and after an item was revealed that did not have a potential contrast, which we will term a non-specific item. These non-specific reveals occurred prior to the specific competitor reveal. For clip art items, there is no increase in length, and for tangram items, there is a numerical but non-significant increase. This suggests that the increase in length is particular to a potential contrast and not a new object in the display. We next assess whether this increase in length could be explained by modification.

Modification

Modification: base rate

Base rate modification was low overall, with 22.4% of the 1671 references having modified referential expressions. Of the 374 modified referential expressions, 62.8% were tangrams, regardless of whether a naming competitor was present in the visual display. Looking at Game 1 only, the modification rate across *all* item types was 28.3%. For Game 2, the modification rate for those items without a naming competitor was 17.4%, but this rose to 54.3% when the items had a potential naming competitor.

Modification: before and after reveal

As stated above, there were three types of items that were occluded: a tangram with an occluded clip-art competitor, a clip-art with an occluded tangram competitor, and an occluded singleton. The proportion of modification for all competitor items *before* its competitor was introduced was 19.6%, whereas the proportion of modification for competitor items *after* its

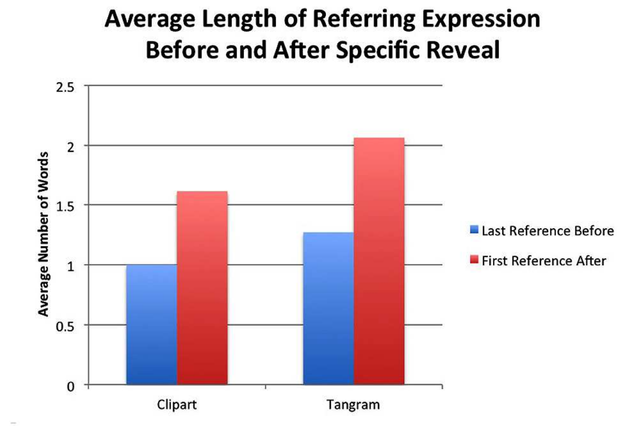


FIGURE 10 | Average number of words in referring expression before and after a specific competitor item was revealed.

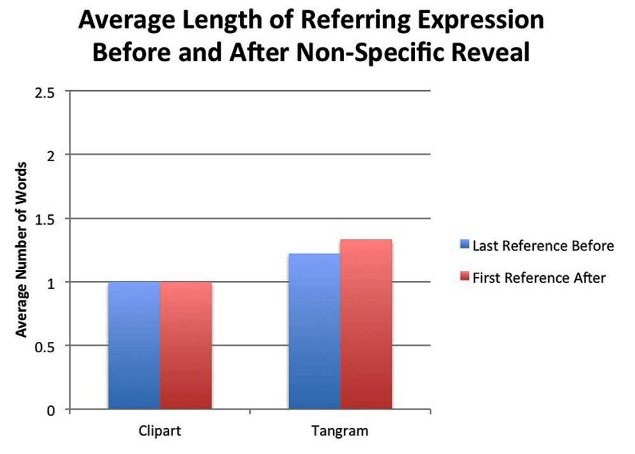


FIGURE 11 | Average number of words in referring expression before and after a non-specific item was revealed.

competitor was introduced was 53.2%. See **Figure 12** for an excerpt of a transcript in which an item previously unmodified is modified after the reveal of the naming competitor. Breaking down the data by item type, tangrams were modified 31.3% before the reveal and 69.4% after the reveal. Clip art images were modified 6.7% before the reveal and 35.8% after the reveal.

The crucial question was whether aspects of the revealed competitor affected entrenchment of a conceptual pact. Looking only at those occluded items with competitors, we used a generalized linear mixed regression model to measure the extent to which item type, number of "before-the-reveal" references, and order of reveal predicts modification. We used modification as the outcome variable and included four predictors: reveal state (i.e., before or after the reveal), item type (tangram or clip art), order of competitor reveal (whether tangram or clip-art was the occluded item), and number of "before" references. A random effect of pair was also included in the model. We introduced both random intercepts and slopes and removed one at a time until the model converged. The model shows that "before-the-reveal" references were less likely to be modified than "after-the reveal" references ($\beta = -2.06$, $SE = 0.43$, $p < 0.001$). In addition, tangrams were more likely to be modified ($\beta = 1.57$, $SE = 0.24$, $p < 0.001$) than clip art images. However, neither the order of the reveal nor the number of "before-the-reveal" references was significant ($\beta = -0.45$, $SE = 0.31$, $p = 0.15$ and $\beta = -0.07$, $SE = 0.09$, $p = 0.48$, respectively). Crucially, then, the pressure to maintain a lexical precedent for the clip art image, which was a better fit to the established name, was not sufficient to lead interlocutors to choose a non-contrastive description for the potential tangram competitor.

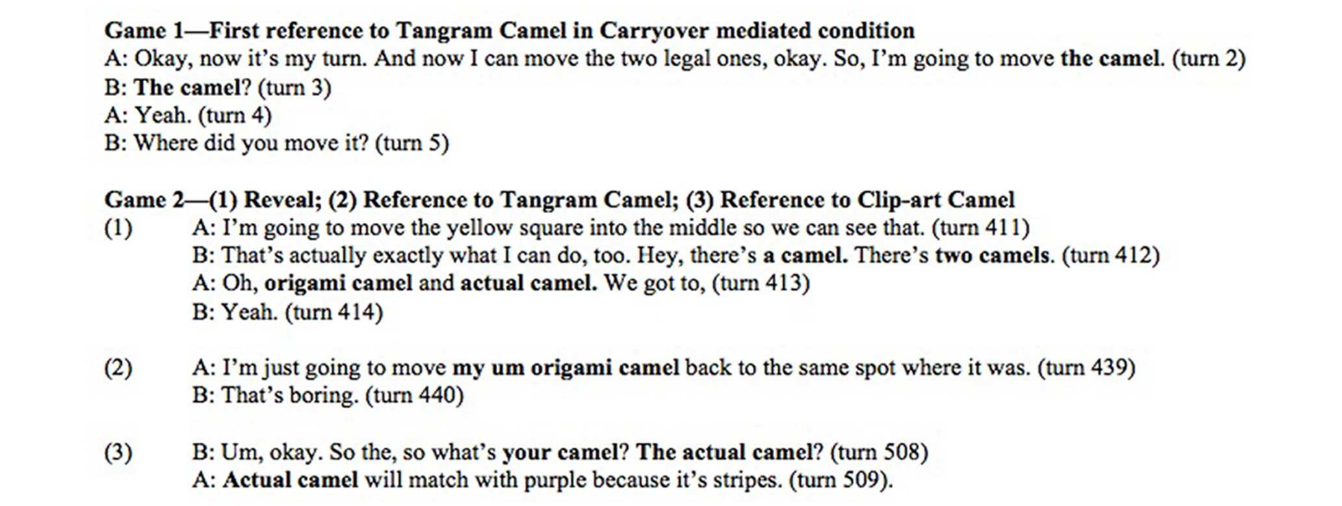
Discussion

The results of Experiment 2 show that conceptual pacts are easily broken when information demands change, even when names have been negotiated and used multiple times. We found clear evidence that names converge and reduce in form with repeated references, replicating previous results. Convergence was noted for both tangram and clip art images suggesting that despite differences in goodness of fit, the naming process was similar. Initial references to tangrams were longer than initial references to clip art images. However, by the third reference, both had converged to the length of a bare noun. The surprising result is that length increases equally for both the tangram objects and the clip art objects after the reveal of a specific competitor, suggesting a breaking of the conceptual pact, even when for a tangram-reveal the speaker might have chosen a name that did not change the name of previously mentioned clip art.

The increase in length was corroborated by an increase in modification. However, there was a higher modification rate overall for tangram objects than clip art. This asymmetry suggests that the images have different goodness of fit with the presumed negotiated name. Accordingly, these images are not obvious contrasts with each other, but could still be conceptualized as such. Although the overall modification rate is higher for tangram objects, both objects are modified after the reveal, and crucially the effect of introducing a potential competitor on modification is not contingent on a particular order of reveal. Thus, modification, here, was tied more closely to competitor presence than the type of competitor. Furthermore, the number of references made prior to the reveal did not predict modification, suggesting that frequency of use of a name is not associated with that name becoming more entrenched. This supports a strongly context-dependent view of conceptual pacts, one in which pacts are flexible enough to accommodate as a new conceptualization of an object, which may occur at any point in the interaction.

GENERAL DISCUSSION

A conceptual pact has historically been framed as a temporary agreement between interlocutors to not only refer to objects by a particular name but also to adopt a particular conceptualization of the referent (Brennan and Clark, 1996). Although Brennan and Clark (1996) emphasized the temporary, context-specific



Game 1—First reference to Tangram Camel in Carryover mediated condition

A: Okay, now it's my turn. And now I can move the two legal ones, okay. So, I'm going to move **the camel.** (turn 2)
B: **The camel**? (turn 3)
A: Yeah. (turn 4)
B: Where did you move it? (turn 5)

Game 2—(1) Reveal; (2) Reference to Tangram Camel; (3) Reference to Clip-art Camel

(1) A: I'm going to move the yellow square into the middle so we can see that. (turn 411)
B: That's actually exactly what I can do, too. Hey, there's **a camel.** There's **two camels**. (turn 412)
A: Oh, **origami camel** and **actual camel.** We got to, (turn 413)
B: Yeah. (turn 414)

(2) A: I'm just going to move **my um origami camel** back to the same spot where it was. (turn 439)
B: That's boring. (turn 440)

(3) B: Um, okay. So the, so what's **your camel**? **The actual camel**? (turn 508)
A: **Actual camel** will match with purple because it's stripes. (turn 509).

FIGURE 12 | Transcript of participants modifying referential expressions for both camels in Game 2, despite using common name in Game 1 (Pair 7).

nature of conceptual pacts, that claim had not been well-established in the literature. In fact the focus on the costs of breaking a lexical precedent suggested that there are strong pressures for interlocutors to maintain a conceptual pacts. To evaluate the context-dependence of conceptual pacts we created conversational interactions in which either the task goals or the potential information demands of the referential domain changed. We asked whether interlocutors would maintain pacts when the local context changes. We found little or no evidence that interlocutors felt bound by names, even negotiated names, when there was a change in these contextual features.

Experiment 1 focused on the extent to which conceptual pacts are bound to specific task goals. We used a puzzle paradigm to elicit unscripted references to objects across different goals with a single partner. Furthermore, we designed the experiment to elicit multiple references to objects in order to assess the possibility of naming changes over time. The variable of interest was not only whether changes occurred but, if so, how they occurred. That is, would a new name for an object require renegotiation? Or would the new context created by the change in the goal structure, from identifying abstract pieces to using those pieces to build an animal, result in abrupt changes in referential descriptions? Indeed, we found that conceptual pacts were easily abandoned when switching from a phase of a game that required identifying single pieces to a phase that involved assembling those pieces into an identifiable animal. Participants were likely to change a name that was already grounded. Even further, they were likely to change that name to an animal-based name abruptly without negotiation. Moreover, interlocutors would return to the previous negotiated name when the action based on the animal-based reference was unsuccessful, i.e., when the parts assumed to be a leg would not fit.

Experiment 2 focused on the degree to which names are entrenched in the face of changing referential alternatives. In this experiment, change was explored more narrowly by measure of length of referential expression and, specifically, modification to an established name. Modification has been shown to be a way of lexically marking contrast in a referential domain (e.g., Nadig and Sedivy, 2002; Sedivy, 2003; Engelhardt et al., 2006; Brown-Schmidt and Tanenhaus, 2006; Brown-Schmidt and Konopka, 2008). And importantly, it connects back to the original work on conceptual pacts (Brennan and Clark, 1996), which used over-modification as a way of determining adherence to a name established in the presence of a contrast set. One example is calling particular footwear in an array "a dress shoe" in the presence of other types of shoes and then maintaining this over-specific form when encountering the object in an array where there are no other shoes. This is taken as evidence that conceptual pacts are maintained (i.e., names are entrenched) despite the change in visual context. However, over-modification is often taken to be more acceptable in conversation as under-modification for various reasons (see e.g., Isaacs and Clark, 1987; Pogue et al., 2016), which reduces the strong claim that name entrenchment and conceptual pacts are creating pressures to use overly specific forms.

Experiment 2 used a targeted language game to create a situation where participants used a name, negotiated or otherwise, and then were presented with an item that may or may not be viewed as a member of a contrast set. This allowed us to examine the degree to which conceptual pacts become strengthened by increased repetition, thereby becoming more resistant to modification. We chose images (i.e., tangrams) that have precedent in language studies of reference and conceptual pacts. However, our modifications to the images allowed us to present a "naming competitor" that is not simply semantically similar but might better fit a common name already grounded. This asymmetry in fit allowed for either option to occur in conceptualization: participants could view them as pairs of a contrast set or choose not to given that one is a better exemplar. Visual similarity helped us avoid objects that might be contrasted by use of a prenominal adjective along a dimension such as size or color. Instead, we saw modifications like "the real camel" and "the origami camel." By occluding some objects including potential competitors, we were able to create conditions where the number of uses of a name, both those that were negotiated and those that were accepted without negotiation, varied as a natural consequence of how the dialog unfolded. As expected, introducing a possible contrast increased modification rates for referential descriptions of objects that had been referred to previously. Somewhat surprisingly, though, was that there was no effect of number of mentions on modification rates. This suggests that even if repeated mention might strengthen a conceptual pact, the effect is not strong enough to affect how a new object will be conceptualized, which would be the case if the conceptual pact were resistant to modification.

Taken together the results suggest that interlocutors choose names that are highly dependent upon local task goals and informational demands, and these names can change rapidly as different aspects of the context change. One the one hand, this is not surprising. Use of referring expressions is highly fluid, as indexed by the many to many mappings that can be observed in repeated references to objects in a discourse. On the other hand, the fact that pacts are so fluid is surprising. We find strikingly little support for any effects of repetition of a name. Once the context changes, the referring expression is determined with respect to that context.

These results suggest that it will be important to embed investigations of reference generation and understanding in richer dialogs where it is possible to investigate the effects of complex goal structures in conversation. Also, incorporating dynamic referential domains in a given interaction that allows for a reconceptualization of alternative objects will help us further understand how goals and naming alternatives influence one another. Targeted language games provide a promising methodological approach for pursuing these investigations.

These results also have implications for computational models of reference generation. For example, existing models have focused on the grounding process (e.g., Heeman and Hirst, 1995), but have not taken into account the temporary nature of the conceptual pacts that are created during grounding. Thus small changes in a local goal might result in abrupt changes to previously grounded expressions. Moreover, evidence for miscommunication might result in returning to referential expressions that were tied to previous goals. There is also a tradition of modeling reference generation and understanding, building on classic work by Dale and Reiter (1995) as an incremental process that takes into account the objects in the referential domain, including the salience of their properties. But what objects are in the referential domain and what properties are salient will be highly fluid and strongly determined by shifting goals.

It is useful here to consider an analogy to work in vision in natural tasks. There is a tradition of modeling shifts in attention to regions of a scene as indexed by using properties such as visual salience derived by integrating multiple feature values at each position within a scene (i.e., saliency maps; see Koch and Ullman, 1985). Moreover, these models correlate with fixation probabilities during viewing of a scene when the observer is not given a particular task. As reviewed by Hayhoe and Ballard (2005, also see Salverda et al., 2011), however, feature-based salience turns out to be a poor predictor of gaze patterns when a participant is engaged in a well-defined task and needs to derive certain information from the visual input to successfully complete the task.

AUTHOR CONTRIBUTIONS

AI and MT together designed the study. AI performed the research and the data analyses. AI and MT together drafted and edited the manuscript.

FUNDING

This research was partially supported by NIH grant HD27206 to MT.

ACKNOWLEDGMENTS

We are grateful to Natalie Fuentes for help in transcribing the data from Experiment 1 and Jenna Duerr for help in running participants and transcribing the data from Experiment 2. Members of the KurTan lab meeting and the Experimental Semantics and Pragmatics (ESP) reading group provided helpful input. We also thank our anonymous reviewers for their helpful feedback.

REFERENCES

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493.

Brown-Schmidt, S., and Konopka, A. E. (2008). Little houses and casas pequeñas: message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition* 109, 274–280. doi: 10.1016/j.cognition.2008.07.011

Brown-Schmidt, S., and Tanenhaus, M. K. (2006). Watching the eyes when talking about size: an investigation of message formulation and utterance planning. *J. Mem. Lang.* 54, 592–609. doi: 10.1016/j.jml.2005.12.008

Brown-Schmidt, S., and Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cogn. Sci.* 32, 643–684. doi: 10.1080/03640210802066816

Carroll, J. M. (1980). Naming and describing in social communication. *Lang. Speech* 23, 309–322.

Carroll, J. M. (1985). *What's in a Name? An Essay in the Psychology of Reference.* New York, NY: Freeman.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Dale, R., and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902_3

Engelhardt, P. E., Bailey, K. G. D., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity. *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Hayhoe, M. M., and Ballard, D. H. (2005). Eye movements in natural behavior. *Trends Cogn. Sci. (Regul. Ed.)* 9, 188–194. doi: 10.1016/j.tics.2005.02.009

Heeman, P. A., and Hirst, G. (1995). Collaborating on referring expressions. *Comput. Linguist.* 21, 351–382.

Isaacs, E., and Clark, H. H. (1987). References in conversation between experts and novices. *J. Exp. Psychol. Gen.* 116, 26–37. doi: 10.1037/0096-3445.116.1.26

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227.

Kronmüller, E., and Barr, D. J. (2015). Referential precedents in spoken language comprehension: a review and meta-analysis. *J. Mem. Lang.* 83, 1–19. doi: 10.1016/j.jml.2015.03.008

Nadig, A., and Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* 13, 329–336. doi: 10.1111/j.0956-7976.2002.00460.x

Paxton, A., Roche, J. M., Ibarra, A., and Tananhaus, M. K. (2014). "Failure to (mis)communicate: linguistic convergence, lexical choice, and communicative success in dyadic problem solving," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, eds P. M. Bello, M. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society).

Pogue, A., Kurumada, C., and Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Front. Psychol.* 6:2035. doi: 10.3389/fpsyg.2015.02035

Roberts, C. (1996). "Information structure in discourse: towards an integrated formal theory of pragmatics," in *Ohio State University Working Papers in Linguistics 49: Papers in Semantics*, eds J. H. Yoon and A. Kathol (Columbus, OH: Ohio State University), 91–136.

Roche, J. M., Paxton, A., Ibarra, A., and Tanenhaus, M. (2013). "From minor mishap to major catastrophe: lexical choice in miscommunication," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, eds M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Austin, TX: Cognitive Science Society).

Salverda, A. P., Brown, M., and Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychol.* 137, 172–180. doi: 10.1016/j.actpsy.2010.09.010

Sedivy, J. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguist. Res.* 32, 3–23. doi: 10.1023/A:1021928914454

Tanenhaus, M. K., and Brown-Schmidt, S. (2008). Language processing in the natural world. *Philos. Trans. R. Soc. B* 363, 1105–1122. doi: 10.1098/rstb.2007.2162

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 31 August 2016
doi: 10.3389/fpsyg.2016.01275

Production of Referring Expressions for an Unknown Audience: A Computational Model of Communal Common Ground

Roman Kutlak, Kees van Deemter and Chris Mellish*

Natural Language Generation Group, Computing Science Department, University of Aberdeen, Aberdeen, UK

This article presents a computational model of the production of referring expressions under uncertainty over the hearer's knowledge. Although situations where the hearer's knowledge is uncertain have seldom been addressed in the computational literature, they are common in ordinary communication, for example when a writer addresses an unknown audience, or when a speaker addresses a stranger. We propose a computational model composed of three complimentary heuristics based on, respectively, an estimation of the recipient's knowledge, an estimation of the extent to which a property is unexpected, and the question of what is the optimum number of properties in a given situation. The model was tested in an experiment with human readers, in which it was compared against the Incremental Algorithm and human-produced descriptions. The results suggest that the new model outperforms the Incremental Algorithm in terms of the proportion of correctly identified entities and in terms of the perceived quality of the generated descriptions.

Keywords: generation of referring expressions, computational model, common ground, audience design, corpus

OPEN ACCESS

Edited by:
Tamara Swaab,
University of California, Davis, USA

Reviewed by:
Melissa Duff,
Vanderbilt University Medical Center, USA
Paul Edward Engelhardt,
University of East Anglia, UK

****Correspondence:***
Kees van Deemter
k.vdeemter@abdn.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *01 September 2015*
Accepted: *10 August 2016*
Published: *31 August 2016*

Citation:
Kutlak R, van Deemter K and Mellish C (2016) Production of Referring Expressions for an Unknown Audience: A Computational Model of Communal Common Ground. Front. Psychol. 7:1275. doi: 10.3389/fpsyg.2016.01275

1. INTRODUCTION

A large body of research in psycholinguistics investigates the extent to which speakers tailor their utterances to their addressees, a phenomenon known as *audience design* (Clark and Murphy, 1982; Clark and Wilkes-Gibbs, 1986b; Isaacs and Clark, 1987; Clark and Brennan, 1991). Referring expressions (henceforth, REs) are a natural focus for research on audience design, because they aim to identify a referent uniquely for an audience; if the RE includes information unknown to the hearer, then the hearer may fail to know what or who the speaker talks about. To borrow an example from Appelt (1985), if I tell you to get off the bus "one stop before I do," then my reference to the bus stop will tend to misfire, because you do not know where I will get off the bus. The link between knowledge and reference makes REs a suitable focus for research on Audience Design. The present article follows this well-trodden path, using computational models, and experiments with human participants. Computational models will be employed because they are the most explicit and detailed models of reference production that are currently on the market (see van Deemter, 2016; also Section 2 below); controlled experiments with human participants will help us ground our computational model in actual human behavior.

Audience design is difficult at the best of times. This article focusses on a class of situations in which the process is complicated further by the fact that the speaker addresses an unknown audience, for example as when a novelist writes a book, or a scientist addresses a conference. In these situations, the speaker/writer does not know exactly who is reading or listening, let alone what the listeners know; moreover, different listeners know different things, hence a RE that works well for one listener might work badly for another. For concreteness, we focus on REs that serve to identify personalities in the public domain (i.e., famous people); generalizations to other publicly known entities—such as companies, towns, sports clubs, and so on—suggest themselves naturally.

Thus, this article presents a computational model of reference to famous people, under uncertainty about the hearer's knowledge. As will be explained, our model rests on three factors. The first is the likelihood that a given property of the referent is known; we call this the **Knowledge** factor. The second is the degree to which a given property is distinctive or useful; we will call this the **Unexpectedness** factor. The third is the completeness of the RE; for reasons that will become clear later, we call this the **Termination** factor. These three factors have never before been combined yet they bear important conceptual similarities to each other. For example, just as it is is important for a speaker to know what her audience knows, it is important to know what information is *useful* to her audience, and what amount of it suffices. In the last analysis, these factors might all be seen as part of what theoreticians have called **Common Ground** (e.g., Clark and Marshall, 1981; Clark, 1996; Beaver, 1997; Vanderschraaf and Sillari, 2009).

In the next section, we review the state of the art in computational models of referring, and the extent to which these models are able to capture the insights that have emerged from psycholinguistic (Section 2). Next, we briefly sketch an elicitation experiment that provides us with a corpus of human-produced REs 3, allowing us to make some initial observations[1]. Our computational model is presented in Section 4; it is experimentally tested in Section 5 and the results are reported in Section 6. The paper concludes with a discussion of the wider implications of our findings (Section 7).

2. COMPUTATIONAL MODELS OF REFERRING AND AUDIENCE DESIGN

Computational models of reference production are also known as referring expression generation (REG) algorithms. Early REG algorithms were, first and foremost, components of dialogue systems (e.g., Winograd, 1972), where they ensure that entities are described in ways that are intelligible to users. Early REG algorithms were not informed by extensive experimentation with human participants. Over the years, however, there has been a gradual shift. First, computational linguists started to incorporate some psycholinguistic findings, hoping this would help them to create more effective referring expressions[2]. Soon after that, REG algorithms started to be tested systematically, for example in terms of the extent to which their output resembles referring expressions produced by human speakers (Passonneau, 1996; Gupta and Stent, 2005; van Deemter et al., 2012). Essentially, this meant that REG algorithms were starting to be seen as *product* models of human behavior, that is, models that focus on the relation between inputs (i.e., the domain and the intended referent) and outputs (i.e., the semantic content of the referring expression), without making further claims about the production *process* (Sun, 2008). Recent REG algorithms are trying hard to simulate human reference production, by modeling phenomena such as variation in language production (Viethen and Dale, 2010; Frank and Goodman, 2012; van Gompel et al., 2012).

In what follows, we first summarize how Audience Design has been understood by theoreticians, and what psychological experiments have taught us about this phenomenon. Next, we discuss to what extent existing REG algorithms address Audience Design. After that, we turn to the challenge outlined in Section 1, namely to model the problem of Audience Design under uncertainty concerning the hearer's knowledge.

2.1. Audience Design in Human Reference Production

Much of our understanding of reference production is based on the idea of Information Sharing (van Deemter, 2016). To convey the idea using a simple example, suppose our shared information is represented in the Knowledge Base of below table. Suppose, furthermore, I have new information for you about the animal *a*: for example, that it is *in a cage*. To communicate this new information to you, I can exploit our shared knowledge, telling you, "the Kenyan lion is in a cage."

Identifier	Species	Origin	Weight	Injuries
a	Lion	Kenya	102 kg	Paws, teeth
b	Lion	China	100 kg	Paws
c	Tiger	China	310 kg	Back

After my utterance, my *privileged* information has shrunk, but our *shared* information has increased because the fact that *a* is in the cage is now part of it. Crucially, a different RE might have been chosen, e.g., "The lion that weighs 102 kg." The choice of referring expression is what REG algorithms are concerned with.

Most authors on REG have written about Information Sharing as if it hinged on what knowledge the speaker knows the hearer to possess. Although this is an important part of it, psychologists, logicians, and game theorists have argued that, strictly speaking, information *p* is only shared between a set of agents (it is also said to be "common knowledge," or "in common ground") if all these agents know *that p and that p is shared*. To borrow an example from Clark and Marshall, suppose I utter the RE "the movie showing at the Roxy tonight" (Clark and Marshall, 1981). If you and I believe this is movie *x*, and I believe that you believe it is *x*, but you believe that I believe it is movie *y* then you will misunderstand me, because you think I'm referring to *y*. A proposition *p* is only shared between you and me if I know that *p*, you know that *p*, I know that you know that *p*, you know that I know that *p*, and so on, using epistemic embeddings of arbitrary depth.

[1] Additionally, as will be explained, this elicitation experiment will play a minor technical role in our modeling of the Termination factor.

[2] For example, the `Incremental Algorithm` of Dale and Reiter (1995) was inspired by Pechmann's findings regarding incrementality (Pechmann, 1989) and by Rosch's work on basic-level values (Rosch, 1978).

Researchers from a number of disciplines have contributed to our understanding of Information Sharing (see Beaver, 1997 for a survey). The philosopher Robert Stalnaker, for example, thought a felicitous utterance should normally fulfill two conditions: it should be consistent with shared information and it should add new information to it (Stalnaker, 1978). This view has sometimes been challenged (e.g., Lewis, 1979 on the notion of *accommodation*), but it matches the relatively simple situations on which REG research has focussed.

To perform information sharing effectively requires the reader to design her referring expressions in a way that allows the hearer to understand what they refer to, a special case of a phenomenon known as Audience Design. Speakers are not always good at Audience Design. The ability in principle of most adult human speakers to reason about "other minds" is well attested, yet speakers and hearers frequently fail to realize exactly what information is shared between them: the extent to which we are able to assess what information is shared is the subject of the so-called *egocentricity* debate. On one side of the debate are psycholinguists who emphasize shared information and its role in communication (e.g., Clark and Wilkes-Gibbs, 1986a; Brennan and Clark, 1996). On the other side are researchers whose experiments have sowed doubt about people's ability to take their knowledge about other minds into account when they speak or listen (Horton and Keysar, 1996; Keysar et al., 2003; Lane et al., 2006), even in situations where it has been made abundantly clear to each interlocutor what the other one knows. Some "doubters" compare our ability to take other minds into account to a fancy espresso machine that you have been given as a present: you own the machine (i.e., you are able to theorize about other minds), yet you may not use it very often (Keysar et al., 2003).

To date, the debate is unresolved, with different researchers attaching different interpretations to experimental results. For example, a study by Wu and Keysar focussed on speakers' choices between names and descriptions (Wu and Keysar, 2007). Participants were shown unfamiliar complex shapes and they were taught equally unfamiliar names for these shapes (e.g., one was called *Abypit*). The authors found that speakers frequently over-use names, tending to produce names where they should have known that the listener had no chance of knowing what they meant. This appeared to confirm the suspicions of the "doubters." However, in a recent follow-up, Heller and colleagues re-examined Wu and Keysar's experiment, and concluded that speakers in that experiment were not over-using names at all: when they used unfamiliar names, this tended to be in situations where sufficient other information was available to permit hearers to know what the name referred to Heller et al. (2012). Essentially, Heller et al. argue, speakers were *teaching* hearers the meanings of the name.

Other publications in this area have given rise to similar discussions, with critics arguing that experimental participants had been put in unusual situations (Brown-Schmidt, 2009). Instead of exploring these issues further, let us see how the reference task can be formulated as part of a computational model.

2.2. Audience Design in Existing REG Algorithms

In accordance with longstanding usage going back to the work of J.S. Mill in the 19th century, we call the set of elements that have a property P the *extension* of P, abbreviated $[\![P]\!]$. Given is a finite domain involving a set M of entities; what entities M contains is shared information between the speaker and the hearer. M contains an element $r \epsilon M$, the target referent. Given are also one or more other elements, the *distractors*, and a set $\mathbf{P}$ of atomic properties, whose extension is shared information between the speaker and the hearer. The REG task may be defined as follows:

> *The* REG *task* If there exists a subset $\{P_1, .., P_n\}$ of $\mathbf{P}$ such that $[\![P_1]\!] \cap ... \cap [\![P_n]\!] = \{r\}$ (so r is the only element in M of which each of these n properties holds true), then REG needs to find such a set. The algorithm needs to make sure that the properties $P_1, ..., P_n$ permit the generation of a RE that is optimally similar to REs produced by human speakers in comparable situations.

Following Dale and Haddock (1991), both the set of properties $\{P_1, .., P_n\}$ and the RE that puts the properties into words is called *distinguishing description*. A distinguishing description is thus a set of properties whose conjunction is true of the referent but not of any other entity in the domain; other entities are called *distractors*. We focus here on algorithms that produce "one-shot" descriptions, that is, which disregard any prior utterances.

Existing REG algorithms have used these ideas in different ways. Some early algorithms generate descriptions that are *minimally* distinguishing (i.e., containing the minimum number of properties), but speakers frequently include additional information (e.g., Levelt, 1989; more recently Arts, 2004; Engelhardt et al., 2006; Koolen et al., 2011). Observations of this kind led to the Incremental Algorithm (IA) (Dale and Reiter, 1995), which assumes the existence an ordered list of attributes, known as a *Preference Order*. The notion of a Preference Order formalizes the idea that some attributes are more likely to be used than others, for instance because they have high utility, or high "codability" (Belke and Meyer, 2002). Color, for example, is thought to have high codability, and this explains the fact that referring expressions frequently contain color in situations where the referent could already be identified by the hearer (i.e., the use of color was logically superfluous). The IA produces different output depending on what Preference Order it uses.

The IA generates a description of a referent r in the following way: The algorithm takes the first attribute from the Preference Order and selects the most attribute that removes the most distractors. If the property rules out one or more distractors, it is added to the referring expression; otherwise it is not added, and the next attribute in the preference order is examined. Crucially, the algorithm terminates when properties $P_1, .., P_n$ have been selected such that $[\![P_1]\!] \cap ... \cap [\![P_n]\!] = \{r\}$. In other words, the algorithm ends *when the algorithms calculates that the hearer is able to identify the referent*, and this is where, arguably, it performs Audience Design. When the algorithm terminates, the description resulting from it is inspected to see whether another property needs to be added: if the description does not contain a property whose attribute is `type`, one such property is added to ensure that the description contains a noun.

2.2.1. Example

Suppose a domain contains three chairs *a*, *b*, *c*, whose color and size are defined. Suppose *a* and *b* are red, while *c* is brown. Furthermore, *a* is large, whereas *b* and *c* are small. Suppose *a* is the referent, and the Preference Order is [color, size]. Then the IA starts examining the most highly preferred attribute, color, selecting *red*. Although this property rules out the distractor *c*, it does not rule out *b*, so the referring expression is not finished yet, and another property needs to be selected. The next property selected is *large*, ruling out the distractor *b*. Both distractors have been ruled out now, and the type of the object is added, that is *chair*. Later processes decide what words to employ for expressing these three concepts, as in "the large red chair." Neither is a *minimally* distinguishing description, since "the large chair" would have been unambiguous.

Dale and Reiter hinted at something like Audience Design (without using the term). Their idea was that when the IA asks whether a property rules out any distractors, the answer is given on the basis of *what the speakers believes to be the hearer's knowledge* about each domain object (Dale and Reiter, 1995, Section 2.3). The IA does not offer a mechanism for assessing the hearer's knowledge. In practice, when implemented, the algorithm invariably uses a simple database of facts (as in our Example above). Clearly, it was not the authors' aim to offer an account of Common Ground.

A model that considers the hearer's knowledge to a slightly greater extent is Horacek (2005). Horacek identifies three types of uncertainty: *knowledge*, *perception capabilities* and *conceptual agreement*. Uncertainty about knowledge occurs when a property may not be known or recognized by the hearer; for example, if the speaker says *"the Basset Hound,"* the hearer may not be able to tell a Basset Hound from other dog breeds. Uncertainty about perception arises, for example, if the hearer does not view a scene from the same position as the speaker, so some properties (e.g., a dog's tail) might be hidden from view. Conceptual agreement uncertainties occur when there is a chance that the speaker conceptualizes a property differently from the hearer; for example, the speaker might describe an object as *turquoise* whereas the hearer would describe it as *blue*.

Horacek (2005) augments the Incremental Algorithm by taking these three types of uncertainty into account. Each property has three probabilities associated with it, one for each of the three types of uncertainty. These three probabilities are combined into one overall probability which helps to determine whether the property in question will become part of the referring expression generated by the algorithm. Although Horacek focussed on very small domains and did not test his algorithm empirically, his work indicates a possible approach to generating definite descriptions under uncertainty. A difficulty is that it is unclear how the necessary probabilities in Horacek's algorithm should be estimated. Our own algorithm (Section 4), by contrast, has computational methods for estimating probabilities at its heart.

2.3. Audience Design in REG: The Challenge Ahead

Considerable effort has been invested in experiments that test the ability of REG algorithms to mimic the REs produced by human speakers (Gatt and Belz, 2010). Almost invariably, these tests have focussed on communicative situations in which it is straightforward to determine what properties are in the Common Ground of the speaker and the hearer. The typical setup of these experiments has been to elicit REs from speakers who are either together in a room with the hearer observing the same visual domain, or they are asked to imagine that they are. Herb Clark described situations of this kind as involving "triple copresence" and observed that these situations make it easy for people to understand what information is in Common Ground.

Although a limited amount of research discusses situations in which the hearer has to make an effort to find out what information is in Common Ground (Garoufi and Koller, 2014a; Paraboni and van Deemter, 2014a), we are unaware of computational models of situations—such as those discussed by Keysar and colleagues—where "egocentric" speakers struggle to realize what is in Common Ground. To mimic speakers' behavior in such situations, a computational model would have to behave as if it has a tenuous grasp of Common Ground, avoiding privileged information in some situations but not in other similar situations; after all, speakers are not unable to understand that the hearer's knowledge differs from their own—they do "get it right" some of the time. To capture this fluctuating behavior, a probabilistic model may have to be designed, which does not always produce the same referring expression in a given situation (Holden and van Orden, 2009; van Deemter, 2016, chapter 6). To do this, however, is not the aim of the present paper.

The problem to which we are about to turn has relevance for the much-debated problem of egocentricity, but instead of examining cleverly designed situations in which speakers *know* very well what the hearer knows but sometimes fail to apply this knowledge, we study the situations discussed in Section 1, where speakers do not have sufficient information to judge what their hearers know. As the domain of our study, we chose the domain of *famous people*. This vast domain forces speaker to guess what properties are likely to be known by hearers. Moreover, the naturalness of the domain allows experimentation with participants without any special training. Finally, this is a domain for which some computational resources exist (such as DBpedia, see Section 4.2) that will prove to be important for the construction of our model.

Our approach owes a debt to Clark and Marshall's insight, that people manage to communicate even in the absence of triple co-presence. These authors suggested that two mechanisms can help us to estimate Common Ground. The first mechanism operates when information is publicly announced, and is called Linguistic Common Ground. For example, when information is broadcast on a train, then this information is accessible to all passengers, so it might be reasonable to assume that it is on Common Ground (barring background noise and lack of attention). The second mechanism operates when people from the same community are exposed to broadly the same sources of information. Residents of Paris, for example, expect other residents to know where the Eiffel Tower is, and they expect other residents to know that they know this, and so on. In connection with situations of this kind, Clark and Marshall coined the term *communal* Common Ground (Clark and Marshall, 1981; Clark, 1996): "common ground based on community membership."

Linguistic and communal Common Ground can only *estimate* Common Ground, because they do not offer an absolute guarantee that each member of the community knows that each member of the community possesses the information involved. Estimation is, accordingly, an important feature of the computational models that will be discussed later in this paper. These models will focus on *communal* Common Ground. The reader may recall that we are focussing on referring expressions that refer in one shot (i.e., disregarding linguistic context). Algorithms that produce REs in linguistic context might be seen as modeling *linguistic* Common Ground; of particular interest in this connection are models in which an algorithm similar to the ones discussed in Section 2.2 are applied to situations in which the shared knowledge base is essentially a piece of text (Siddharthan and Copestake, 2004).

3. INITIAL EXPLORATION: ELICITING A CORPUS OF REs

To gain an initial insight in the descriptions produced by human speakers, we elicited a corpus of descriptions of famous people (Kutlak et al., 2011), using Amazon Mechanical Turk (MTurk), a platform where tasks can be posted that are completed by volunteers for a small financial reward. Participants were told about a game where a speaker produces descriptions of a famous person and a hearer has to guess the name of the person described. Participants were told that the hearer has one attempt to guess the name of the person described and the descriptions produced should help the hearer to identify the described person. Participants were then presented with names of famous people and primed to produce definite descriptions by completing the sentence, "This person was the…". **Table 1** shows some of the 215 descriptions produced by 29 participants. Participants were self-identified native English speakers whose registered address was in the US or the UK (see Kutlak et al., 2011 for more details).

Informal analysis of the corpus suggested to us that many descriptions of the same person share a common core of properties. For example, all descriptions of *Edison* said he invented the light bulb; half of the descriptions of *Hillary Clinton* mentioned that she was a former First Lady. Far from being idiosyncratic, the bulk of the properties employed seemed to be ones that are widely known. This observation set us on a path to designing a **knowledge heuristic**, which estimates what information people are likely to know (Section 4.1).

Secondly, many of the properties mentioned in a description were quite unusual with respect to the general population. Examples of such properties are *the inventor of the telephone* and *received a Nobel Prize*. This is to be expected, because participants were asked to produce descriptions that allow hearers to guess the name of the person described. However, it also suggests that if one wants to simulate human behavior, one might take into consideration how unusual a property is. In other words, although properties should be widely known, they should also be unusual or unexpected. This suggests a second heuristic, for which we use the term **unexpectedness**. In other words, we hypothesized that descriptions should avoid properties that are unexpected yet little known (e.g., *the Warden of the British Royal Mint in 1696*[3]), but also properties that are widely known but too common to be useful (e.g., *the person who had arms and legs…*).

Finally, many descriptions in the corpus were multiply overspecified in the sense that they contained multiple properties that could be omitted without stopping the description from being distinguishing. For example, the description *"This person was the author of The Old Man and the Sea, The Sun Also Rises, For Whom the Bell Tolls, and other famous novels"* contains three properties, each of which identifies *Hemingway* uniquely. Perhaps because participants did not know what the hearers knew, they included extra properties to increase the chance that the hearer identify the referent. To produce a minimally distinguishing description would be to gamble, but it is not obvious *how much* overspecification is required. We use the term **termination** to refer to the factor determining how long the REG algorithm should continue adding properties to the description.

4. THE COMPUTATIONAL MODEL

Reflecting the insights of the previous section, our computational model of reference production is composed of three heuristics, each of which corresponds to one of the factors discussed above.

[3] Isaac Newton.

TABLE 1 | A sample of referent names and corresponding descriptions in the corpus.

Name	Description
Albert Einstein	This person was the author of the theory of relativity
	This person was the physicist who developed the Theory of Relativity that revolutionized how we understand space, time, and gravity
	This person was the German-American mathematician
Thomas Edison	This person invented the light bulb
	This person was most famous for his inventions of the light blub and the phonograph
	This person was the inventor of the light bulb, phonograph, and movie projector
Elvis Presley	This person was the King of Rock "n" Roll
	This person was the King of Rock and Roll, born in Tupelo, Mississippi, who had Graceland built
	This person was one of the most popular singers ever, with hits including Blue Suede Shoes and Jailhouse Rock

This heuristic-based approach makes the model transparent, because one can see what each heuristic contributes (though of course we have no evidence that they have separate reality in the human mind). Having individual heuristics also makes the model more extensible, because new heuristics can easily be added. For example, the model could be extended by a heuristic that takes into account the context in which the description appears.

A general remark about the way in which our three heuristics were developed is in order. There are many ways in which the loosely explained ideas of the previous section may be turned into precisely defined metrics that can be applied to actual data. The danger of testing hypotheses involving a large number of metrics is that the risk of type I errors (false positives) is considerable. Our solution to this problem was to use a two-stage approach. During the first stage, we implemented a number of metrics that appeared to be plausible on the basis of earlier work, and we did a pilot test on all of them. During the second stage, those metric that performed best during the pilot were tested again, using a new set of stimuli. Any metrics that achieved a good performance by a chance during the first stage are likely to fail during the second. This approach allowed us to test a large number of options while keeping the risk of type I errors low.

4.1. Knowledge Heuristic

In order to generate useful descriptions of people, the computational model should select properties that are known by the hearers, and this involves estimating hearers' knowledge. We take as our starting point the idea that speakers use community membership to estimate what hearers know (Clark and Marshall, 1981). Experimental evidence shows that speakers are often able to distinguish between knowledge that is available to members of specific communities or to outsiders. For example, Krauss and Fussell (1991) cite an experiment by Kingsbury (1968), who asked random pedestrians in Boston for directions to a local department store. Kingsbury asked one third of his subjects (a) "Can you tell me how to get to Jordan-Marsh?" using a local dialect, one third (b) "I'm from out of town. Can you tell me how to get to Jordan-Marsh?" using the same dialect, and one third (c) "Can you tell me how to get to Jordan-Marsh?" using his native rural Missouri—a dialect not often heard in Boston. Respondents in groups (b) and (c) provided longer and more detailed responses than in group (a). Related conclusions can be drawn from Bromme et al. (2001) and Nickerson et al. (1987).

Our hypothesis is that the knowledge of a community can be estimated by examining documents produced by the community: the more frequently a fact is mentioned, the more likely are the members of the community to know this fact. This may be motivated by two considerations: firstly, an author who reports a fact will tend to know this fact; secondly, if a fact is recorded frequently, then it may be read often, making it more likely to be remembered (Atkinson and Shiffrin, 1968).

Given our hypothesis, two additional things are required: (a) a corpus of documents that represents the target community, and (b) a metric that allows us to calculate how likely it is that addressees know a given fact. As our corpus, we used the World Wide Web, and to gain information from it we used the search engine Google. The World Wide Web has been successfully used as a corpus before (e.g., Turney, 2001; Keller and Lapata, 2003). The advantage of using a search engine such as Google is its ability to take synonyms and morphological variations into account and to ignore irrelevant words that separate the search terms. Since our queries will use English search terms, the documents retrieved by the search engine are also in English, so we hypothesize that they represent the community of those English speakers who regularly access the World Wide Web.

To implement the Knowledge Heuristic, we experimented with a number of computational metrics of co-occurrence based on the counts of documents containing certain facts, or metrics based on probabilities derived from these counts. Below, we list the four metrics that performed best in our pilot ("stage 1") experiment, all of which are existing metrics for measuring the strength of association between words. Each of the metrics assumes that a *context* for the words has been defined. These contexts are often defined as a limited number of words before or after the target word or a short frame such as a paragraph in which the target word occurs, but this is not suitable for our purpose, because a fact about a person can be mentioned further away from the person's name, especially if the name is pronominalized in consequent paragraphs. Therefore, we used as our context the entire page returned by the search engine.

4.1.1. Frequency

The simplest measure of association between a person and a property (a fact about a person) is the frequency of occurrence of the name and the property together in a corpus. Taking a collection of documents as a corpus, frequency corresponds to the count of articles that contain the name and the property. This association is then the value of $count(n, p)$ where n stands for the name of an entity and p is the property in question.

4.1.2. Conditional Probability

More sophisticated measures are conditional probabilities as in Equations (1) and (2), where Equation (1) measures the probability of occurrence of the name given that a property occurs, and Equation (2) measures the probability of occurrence of a property given that the name of the person occurs. While the former measure normalizes the results by the frequency of the property, the later measure takes into account how famous each person is.

$$assoc_{prob}(n, p) = P(n|p) = \frac{count(n, p)}{count(p)} \tag{1}$$

$$assoc_{prob}(p, n) = P(p|n) = \frac{count(p, n)}{count(n)} \tag{2}$$

4.1.3. Pointwise Mutual Information

(PMI) Fano (1961) compares how often two events x and y occur together. PMI exploits the fact that if two terms appear together often, their joint probability ($P(n, p)$) will be higher than if they were independent ($P(n)P(p)$). The value of PMI is positive for terms that co-occur and negative otherwise.

$$assoc_{PMI}(n, p) = log_2 \frac{P(n, p)}{P(n)P(p)} \tag{3}$$

- Albert Einstein
- Bill Gates
- Christopher Columbus
- Elvis Presley
- John F. Kennedy
- Julia Roberts
- Marilyn Monroe
- Princess Diana
- Sigmund Freud
- William Shakespeare

FIGURE 1 | Names of famous people used in the pilot test of the potential metrics for the Knowledge Heuristic.

A problem with PMI is that infrequent words that only appear together achieve a disproportionately high score. This is undesirable, because in order for a property to be in common ground, it also has to be frequently mentioned. To mitigate this type of problem, Hodges et al. (1996) suggest multiplying each PMI score by $count(n,p)$. To balance out the large differences between the frequencies of frequent items (hundreds of thousands of results) and less frequent items (hundreds of results), we multiply the PMI score by the square root of the count. The final formula used for calculating the association is as in Equation (4).

$$assoc_{PMI}(n,p) = \sqrt{count(n,p)} * log_2 \frac{P(n,p)}{P(n)P(p)} \qquad (4)$$

These four metrics were first tested in a pilot experiment and the best performing metric was then re-tested (in what we call the "main experiment" in this section) using a different set of stimuli.

Given that the setup of the pilot and the main experiment was essentially identical, we describe the the method and the procedure only once. Participants, materials, and results are reported separately for both the pilot and the main experiment.

4.1.4. Materials and Method of the Pilot Experiment

For the pilot experiment, we selected 10 people, each of whom was famous enough that his/her name occurred on the BBC Historical Figures page[4]. The 10 people were selected in such a way that they varied in terms of how well known they were likely to be. The names of the selected people are listed in **Figure 1**.

For each referent we selected information from Wikipedia and the BBC Historical Figures pages. We used our own judgment (informed by the frequencies of properties from the corpus described in the previous section) to select properties that covered a range of likelihoods of being known. That is, for each referent, we selected a number of properties that were likely to be known by anyone who knows the referent (e.g., the referent's occupation), and properties that were likely to be known only by people who have a more detailed knowledge of the referent.

For each referent, we included 5 filler properties that were not true of the referent. Each trial contained 5 true properties and 5 filler properties for each person, presented in randomized order. This resulted in a total of 100 statements (10 referents, 10 properties each). To keep the task manageable, statements were randomly split into 5 groups of 20 statements.

TABLE 2 | List of properties true of Albert Einstein.

Property	Percentage	Frequency	Rank
Albert Einstein was a physicist	80.95	827000	4.00
Albert Einstein invented the theory of relativity	80.43	69600	3.00
Albert Einstein was German	67.39	1060000	5.00
Albert Einstein emigrated to the United States	47.06	20200	2.00
Albert Einstein was a professor at the Karl-Ferdinand University in Prague	30.30	652	1.00

The percentage of affirmative answers show what percentage of participants believed the statement to be true. Rank shows how the corresponding properties ranked according to the Knowledge Heuristic. Spearman correlation between percentage and frequency $r_{s(48)} = 0.67; p < 0.001$.

4.1.5. Participants and Procedure of the Pilot Experiment

The pilot experiment was conducted online using Amazon Mechanical Turk (MTurk) to ensure a large number of participants. A total of 755 MTurk users started the experiment, but only 216 completed the experiment[5]. A further 12 were removed because the number of errors (counted as answering "true" to a false statement or vice versa) was higher than 5 (avg. + SD). The resulting dataset contained 4080 answers produced by 204 participants (78 female, 126 male).

The first screen showed instructions on how to answer and how to navigate the website and also urged the participants to rely on their own knowledge and avoid using external resources to answer the questions. The participants were then asked to fill in some information such as their sex, age group and interests. The participants were then randomly assigned to one of the groups. The participants viewed one statement at a time and were asked to select one of the options (`true`, `don't know`, `false`). Participants could also add a comment to each statement; at the end of the experiment they were given an opportunity to offer additional open comments.

4.1.6. Results of the Pilot Experiment

The responses from participants were aggregated per statement, resulting in each statement having a percentage of affirmative answers (answers where participants selected `true`). Only true statements were analyzed. **Table 2** shows example statements and their scores for Albert Einstein.

Table 3 shows correlations between the metrics and the percentages of affirmative answers calculated from participants' answers. Frequency, $p|n$ and PMI performed very well. Given the simplicity and the performance of the Frequency metric, we decided to chose this metric for the Knowledge Heuristic, where its performance in conjunction with the other metrics was to be tested.

4.1.7. Materials of the Main Experiment

Similarly to the pilot experiment, we selected 10 famous figures (see **Figure 2**), each of whom was famous enough that their

[4] http://www.bbc.co.uk/history/historic_figures/

[5] Investigation of the large number of incomplete answers and potential issues of data quality are addressed in Section 4.1.8.

TABLE 3 | Pilot results: Spearman correlation between the metrics and knowledge of hearers.

	Frequency	*n*|*p*	*p*|*n*	PMI
$r_{s(48)}$	0.667	−0.063	0.672	0.628
p-value	0.000	0.663	0.000	0.000

- Admiral Nelson
- Alfred Nobel
- Andy Warhol
- Duke of Wellington
- Emperor Hirohito
- Ernest Hemingway
- Florence Nightingale
- Heinrich Himmler
- Louis Pasteur
- Plato

FIGURE 2 | Names of famous people used in the test of the Knowledge Heuristic.

names occurred on the BBC Historical Figures page[6]. Each trial contained 7 true properties and 5 filler properties for each person, presented in randomized order. This resulted in a total of 120 statements (10 referents, 12 properties each). The statements were also randomly split into 5 groups of 24 statements (14 true, 10 false in each group).

4.1.8. Participants of the Main Experiment

The main experiment involving the Knowledge Heuristic was likewise conducted online using Amazon Mechanical Turk (MTurk). The pilot experiment had a large proportion of users from India. Given that the experiment required knowledge of "western" culture, we decided to limit the main experiment to US and UK MTurkers. Furthermore, participants also had to successfully pass a cloze test (Stubbs and Tucker, 1974), guaranteeing that only highly fluent speakers of English would pass.

The main experiment was undertaken by 71 English speakers. 5 were discarded because they did not finish the experiment and a further 5 participants were removed because the number of errors they made was more than 4 (mean + 2 * SD). The total number of participants was 61; of these, 30 females, 29 males and 2 unspecified.

4.1.9. Results and Discussion of the Main Experiment

Answers were aggregated by statement and the resulting percentages were correlated with scores assigned by the Knowledge Heuristic. **Table 4** shows examples of the statements used in the experiment, along with the percentages of answers where participants selected `true` and the ranks assigned by the Knowledge Heuristic. The search queries were run in June 2014. We found a high correlation between the estimates produced by the heuristic and the percentage of people who knew given facts [$r_{s(68)} = 0.73; p < 0.001$]. The heuristic performs very well if we compare the results of the heuristic with the correlation of estimated and actual knowledge of human speakers as reported in Fussell and Krauss (1992). In their experiments, Fussell and Krauss report that the average correlation of people's estimate of knowledge of others and their actual knowledge was 0.67 (note that this was a Pearson correlation and our results report a Spearman correlation).

TABLE 4 | List of properties of Ernest Hemingway.

Property	Condition	Percentage	Rank
Ernest Hemingway was a writer	True	100.0	1
Ernest Hemingway was American	True	100.0	2
Ernest Hemingway received the Nobel Prize in Literature	True	63.6	5
Ernest Hemingway is the author of For whom the bell tolls	True	54.5	4
Ernest Hemingway committed a suicide	True	50.0	3
Ernest Hemingway was British	False	27.3	–
Ernest Hemingway was born in Oak Park	True	25.0	6
Ernest Hemingway received the Italian Silver Medal of Bravery	True	20.0	7
Ernest Hemingway is the author of A tale of two cities	False	13.3	–
Ernest Hemingway invented dynamite	False	0.0	–
Ernest Hemingway died in a plane crash	False	0.0	–
Ernest Hemingway was born in Paris	False	0.0	–

Condition shows whether a property was true or false (a filler) and the percentage of affirmative answers shows what percentage of participants believed the statement to be true. Rank shows how the corresponding properties ranked according to the Knowledge Heuristic. Spearman correlation between percentage and rank $r_{s(68)} = 0.73; p < 0.001$.

These results suggested to us that our Knowledge Heuristic is a viable starting point for a computational model of people's knowledge. Note that, strictly speaking, our heuristic was not tested in terms of its ability to capture Common Ground. After all, for a proposition to be in Common Ground (in the strict sense) between all members of a community, it is not sufficient for the proposition to be known by all members: additionally, all members should know that the proposition is in Common Ground; the recursion in this formulation implies an infinite sequence of epistemic iterations (Section 3). Testing our heuristic's ability to capture classic Common Ground would have been very difficult, which is why we settled for a simpler test. Whether this leads to a heuristic that is useful for our purposes is something we were only able to determine when our complete model was tested (Section 5).

However, the Knowledge Heuristic is not sufficient for producing referring expressions, because it does not take into account whether a piece of information is distinguishing. For example, if the heuristic had to decide between properties such as *X is a scientist* and *X is a physicist*, it would assign a higher score to the former. The next section will discuss a heuristic that will balance this deficit.

4.2. Unexpectedness Heuristic

The analysis of the corpus showed that some of the properties selected by human speakers are unexpected with respect to the population as a whole (e.g., "*inventor of dynamite,*" "*received a Nobel prize*"). In order to tell the unexpected properties apart from the more common ones, it would be instructive to look

[6] http://www.bbc.co.uk/history/historic_figures/

at DBpedia. DBpedia is an ontology derived from Wikipedia, a free encyclopedia created by the community of its users. DBpedia extracts some of the information available as free text in Wikipedia and encodes it in machine-readable form. The data is ontologically structured, for example `Physicist` is a subclass of `Scientist`, and `Scientist` is a subclass of `Person`; this information would be difficult to infer from free text. DBpedia suits REG algorithms as it describes each entity by means of properties in the form ⟨*attribute*:*value*⟩. Furthermore, each entity in DBpedia has a type (e.g., ⟨*type* : *person*⟩ or ⟨*type* : *scientist*⟩), which is something many REG algorithms require, as we have seen in Section 2.2.

Finding the right unexpectedness heuristic proved to be challenging. We are interested in properties that are unexpected for our audience. For example, being awarded the Nobel prize is unexpected because only a handful of people receive this prize every year; on the other hand, having a mother is expected as everyone has a mother. Some REG algorithms achieve unexpectedness by selecting properties that are highly discriminating, as defined by Dale's Discriminatory Power (DP; Dale, 1992). The DP of a property is defined as $(N - n)/(N - 1)$ where N is the total number of entities and n the number of entities with a given property. The result of the function takes values between 0 (the property is true for all entities in the context, hence it is not discriminating at all) and 1 (the property is true for only 1 entity in the context, hence it is highly discriminating).

As we are aiming for interesting properties, rather than distinguishing one, DP seemed to be less suitable. One of the problems with DP is the uniform treatment of properties across entity types. For example, in the case of our people domain, a property such as ⟨*Nobelprize* : *literature*⟩ has almost exactly the same score for a writer and a physicist, which is undesirable. A second problem with DP relates to the context we used it in. Although DBpedia contains millions of entities, the properties are very sparse, therefore DP scores many properties very highly: the DP tends to place many properties close to 1; as a result, it does not provide a lot of information about these properties.

One field that studies the recognition of unusual patterns is data mining. Geng and Hamilton (2006) performed an extensive survey of statistical measures of interestingness and categorized them into concepts such as *surprisingness, peculiarity, utility, etc.* Surprisingness or unexpectedness was typically defined in terms of contradicting a person's existing knowledge or expectations; formalizations of this idea often make use of conditional probability; for example, winning a Nobel Prize may be unexpected (even) for a physicist because the probability P(Nobel | Physicist) is low.

To test the ability of a metric to measure unexpectedness, we conducted an experiment similar to the one on the *quality* of expressions in Section 5.3. Participants were told: *"Imagine your friend tells you he has read something interesting about a person. He tells you the name of the person, but you've never come across the name, so you ask who this is. Your friend wonders what to tell you about this person. Please rate, for each of the facts below, how interesting you would find this fact. Please rate each fact individually (i.e., in isolation from the other facts in the list)."* We tested over 30 statistical measures from Geng and Hamilton (2006) but our pilot found no reliable correlation between the predictions of a metric and people's judgements.

While we were unable to find a metric that performed well on its own, we were able to use our experience with the 30 existing metrics to construct a new metric that showed good results when combined with the Knowledge Heuristic. The metric described by Equation (5) has a greater range of values than DP, as can be seen from **Table 5**, assigning higher scores than DP to properties that are less frequent. For instance, in **Table 5**, ⟨`type:astronaut`⟩ is much more unexpected than ⟨`type:scientist`⟩ according to this metric than according to DP, which we believe to be as it should be (i.e., intuitively, astronaut is a more interesting property than scientist). Unlike DP, our formula looks at a property B of a referent in connection with the *type* of the referent, A.

$$Score_{unexpectedness}(A, B) = \frac{P(AB)/P(A)P(B)}{\sqrt{P(A)P(B)P(\neg A)P(\neg B)}} \qquad (5)$$

4.3. Termination Heuristic

The last heuristic determines how much information should be included in a description. Like most REG models (Section 2.2), our algorithm will add properties one by one. Consequently, the number of properties in the description generated depends on when the algorithm terminates. For this reason, we refer to this as *termination*. As we have seen in the corpus, human-produced descriptions often contain several properties that are uniquely distinguishing on their own. Using the traditional approach of terminating when all distractors are ruled out would never produce such descriptions.

As with other heuristics, we tested several methods for terminating the algorithm. Assuming that documents produced by a community can tell us something about the knowledge of the community, we focused on document-retrieval based methods. The intuition was that retrieving documents that contain properties listed in a description can tell us something about the distractors that the audience is likely to be aware of. For example, if our description contained the properties *"singer"* and *"rock 'n' roll,"* a large number of documents would match this description. If the description also contained the property *"singer of Jailhouse Rock,"* the set of matching documents would be much smaller. Three methods were tested. Each time a property was added to a description, the Google search engine was queried using the new description; the search engine returned the number of results plus snippets of text from each document retrieved.

TABLE 5 | Unexpectedness of some properties, by Equation (5) and by DP, calculated across DBpedia.

Property	Unexpectedness (Equation 5)	Discriminatory Power
⟨type:thing⟩	0	0.0
⟨type:person⟩	11	0.6233
⟨type:scientist⟩	89	0.9962
⟨type:astronaut⟩	430	0.9998

The first method is based on the frequency with which the name of the referent occurs in the documents retrieved when a given description is used as a query. The algorithm constructs a description by adding properties to it. At each stage, the method focusses on the description under construction and examines the snippets returned by the search engine (given that the description is used as a query) and counts what percentage of the snippets contain the name of the intended referent. If this percentage crosses a threshold, the algorithm terminates; otherwise, a further property is added to the description. The process repeats until the threshold is crossed or the description contains the maximum number of properties. Based on our analysis of the corpus in Section 3, the maximum number we allowed was 7 (average + 2 SD).

The second method uses only the counts of the results (i.e., documents returned) that contain the name of the target referent; let's call this number *N*. As in the first method, every time the algorithm adds a new property to the description, the search engine is queried using that description. The number of results will decrease, since fewer and fewer documents match the query. As soon as the number of results falls below a pre-determined fraction of *N* (or when it reaches the maximum number of properties), the algorithm terminates.

The third method used the *differences* between the numbers of retrieved documents as a description is being created (i.e., the slope of the graph). Initially, every addition of a property results in a large reduction in the number of results, but as the number of properties in a description increases, the reduction becomes smaller. The heuristic uses this observation to decide whether a description contains enough properties. As soon as the addition of a property does not result in a large difference in the number of matching documents, the heuristic terminates the algorithm. The meaning of "large" is determined by a predetermined threshold. The full heuristic is described by **Figure 3**.

All thresholds were based on the counts of documents containing the name of the referent and an empirically determined coefficient. The reason for using the number of documents containing the referent is that the amount of content needed to describe a person is related to how well known the person is. It seems plausible that very famous people require fewer properties to identify them than less known people will require. The coefficients were derived from a subset of the corpus described in Section 3 annotated with semantic properties.

The three methods were tested using the corpus of Section 3. Each method took as an input the name of the referent and a list of properties that human participants had written to describe the referent, ordered from most to least frequent. The length of the description produced by each method was compared against the average length of all descriptions of a given referent. The score of the description created by the heuristic was calculated using (Equation 6), a standard *z*-score, where μ_i is the average length of the descriptions of *i*, and σ_i is their standard deviation:

$$score(description_i) = \frac{|\mu_i - length(description_i)|}{\sigma_i} \qquad (6)$$

```
Termination Heuristic
  Variables and functions:
  $D$ - description
  $r$ - referent
  $x$ - parameter for setting the threshold
  $name(r)$ - the name of the target entity
  $count(q)$ - the number of documents matching a query
  $lastCount$ - the last number of documents matching a query
  $MaxLength$ - the maximum length of a description

function SHOULDTERMINATE(D, r, x, lastCount)
    $N \leftarrow count(name(r))$
    if $D \neq \{\}$ then
        $difference \leftarrow |lastCount - count(D)|$
        if $(difference < N/x)$ OR $length(D) == MaxLength$ then
            return true
        else
            $lastCount \leftarrow count(D)$
            return false
        end if
    else
        return false
    end if
end function
```

FIGURE 3 | Pseudocode describing the termination heuristic. The heuristic returns `true` (and terminates the algorithm) when adding a property to a description does not result in a large decrease in the number of matching documents.

The third method produced the best results, with an average score of 0.98. The average number of properties per description produced by people was 3.349 with *SD* = 1.754 and the average number of properties selected by the termination heuristic was 3.41 with *SD* = 1.34. This heuristic was selected for the final computational model. The heuristic is described using pseudocode in **Figure 3**.

4.4. Combining the Three Heuristics

The above heuristics were combined as in **Figure 4**. The algorithms start by calling the function `MakeReferringExpression` and passing as parameters the referent and a list of properties true of the referent. An initially empty list *D* is created, which will later contain the properties used in a description. A score is assigned to each property, based on the combination of the Knowledge Heuristic and the Unexpectedness Heuristic. A loop is then entered where the property with the highest score is taken and removed from the list. Next, it is checked whether the algorithm should terminate. If the Termination Heuristic returns `true`, the algorithm returns the list *D* as the final description. If the Termination Heuristic returns false, the algorithm adds the current property into the list *D* and loops back to selecting the next property with the highest score. The loop repeats until the Termination Heuristic stops the algorithm.

The score for each property is calculated as follows. Each property is tested by the Knowledge Heuristic and the Unexpectedness Heuristic. The scores assigned by each heuristic are scaled to range between [0–1] using the function *Scale*. The reason for scaling the values is that each heuristic is using a different scale. (Taking an average would not make sense if, for example, the Knowledge Heuristic produced values between 0 and 10^{12} and the Unexpectedness Heuristic would produce

```
Algorithms 1 and 2
Variables and functions:
$r$ – a referent
$D$ – a description
$P$ – a list of properties true of $r$
$name(r)$ – returns the name of the referent $r$
$type(r)$ – returns the type of the referent $r$
$top(S)$ – returns the element of a set of pairs with the largest second element
$first(pair)$ – returns the first element of a pair $pair$
$Score(p, r, P)$ – returns the score assigned to property $p \in P$
$ShouldTerminate(D, r, x)$ – function returns true if the algorithm should terminate
$Score_{KH}$ – score following the Knowledge Heuristic.
$Score_{UH}$ – score following the Unexpectedness Heuristic.

function MakeReferringExpression($r, P$)
    $D \leftarrow \{\}$
    $scores \leftarrow \{(p_i, Score(p_i, r, P)) | p_i \in P\}$
    while $scores \neq \{\}$ do
        $p_i \leftarrow first(top(scores))$
        $scores \leftarrow scores - p_i$
        if $ShouldTerminate(D, r, x)$ then
            return $D$
        else
            $D \leftarrow D \cup \{p_i\}$
        end if
    end while
end function

function Score($p, r, P$)
    $scores_{KH} \leftarrow \{Score_{KH}(name(r), p_i) | p_i \in P\}$
    $s_1 = Scale(Score_{KH}(p), min(scores_{KH}), max(scores_{KH}))$
    $scores_{UH} \leftarrow \{Score_{UH}(type(r), p_i) | p_i \in P\}$
    $s_2 = Scale(Score_{UH}(p), min(scores_{UH}), max(scores_{UH}))$
    $score \leftarrow (s_1 + s_2)/2$
    return $score$
end function

function Scale($x, min, max$)
    if $(max - min) \neq 0$ then
        return $\frac{x-min}{max-min}$
    else
        return $0.5$
    end if
end function
```

FIGURE 4 | Pseudocode for algorithms Alg1 and Alg2 .

scores between 0 and 1). The function for calculating the score of a property is described in **Figure 4**.

The algorithm is similar in some ways to the Greedy Heuristic (Dale, 1989) because it always chooses the "best" property (based on the three heuristics). Unlike the Greedy Heuristic, our algorithm does not re-calculate the scores of the properties after each iteration. In this regard, the algorithm is more similar to the IA because once the properties are ordered by the heuristics, the order does not change. Unlike the IA, however, the preference is computed for each referent individually.

We also tested a baseline heuristic for terminating the algorithm based on the average number of properties in the human produced descriptions, which is 3. This baseline is not sensitive to the content of the description, risking descriptions that are too general to allow identification. Our experiment in Section 5 therefore tests two different versions of the algorithm: both used the Knowledge Heuristic and Unexpectedness Heuristic to rank the properties but Alg1 used always 3 properties per description whereas Alg2 used the document-retrieval based termination heuristic.

5. EVALUATING THE MODEL

Evaluation of the model focused on three aspects of the descriptions produced. We decided that the main aspect to focus on was the number of successfully identified referents, because identification is the main purpose of referring. The second aspect was *naturalness*, defined by the statement: *"How natural does the description read to you? (For example, could one of your friends produce such a description?)"*. This should tell us something about the human-likeness of the descriptions produced. After all, a description may be effective yet unlike anything that a human speaker is likely to produce. The third aspect was *quality*, defined using the statement: *"Suppose you did not know this person, how good would you find the description? (Does it give a good idea* *of what sort of person it is or was?)"*. We felt it important to assess this because an addressee may not know the referent, in which case the number of successfully identified referents (our first metric) misses the point.

5.1. Algorithms/Models Considered during Evaluation

Computational algorithms were tested along with two types of human-produced descriptions. The first were short descriptions available in DBpedia. Most entities in DBpedia contain the property description, which is comparable to the first line in a Wikipedia article describing an entity. Where the description was not available in DBpedia, we used the first line in Wikipedia (not counting the dates of birth and death).

The second type of human-produced descriptions were created specifically by a native English speaker. A postgraduate student with experience in natural language processing was given a set of 100 names and asked to create English descriptions matching the scenario presented to the participants. That is, the student was asked to produce descriptions of the names so that other US/UK people can guess who the described person is. The student was not aware of the aims of our research but had access to external resources (e.g., the World Wide Web), to ensure that he was able to create descriptions for all entities and not only the ones known to him. We will talk about the "algorithm" DBP when referring to DBpedia descriptions and about the "algorithm" Human when referring to the descriptions produced by the student.

The IA is often used as a reference point against which other algorithms are compared. However, the performance of the IA depends crucially on the chosen preference order. To find a good preference order, we used the semantically annotated part of the corpus described in Section 3. The annotation is similar to the annotation of the TUNA corpus (van Deemter et al., 2012). The preference order was found by taking the annotated properties and ordering them by their frequency from the most to the least frequent attribute. This method has been used by a number of researchers (e.g., Koolen et al., 2012; van Deemter et al., 2012). The first 10 attributes of the preference order were *type, occupation, nationality, country, starring, author, known for, genre, gender, death cause*.

Given the above, we set out to compare 5 classes of descriptions: the ones generated by the algorithms Alg1 and Alg2 the one generated by IA (with the preference order as stipulated), and finally the two human-produced descriptions DBP and Human.

"Descriptions" produced by a computer algorithm are nothing more than a list of properties. To allow participants to judge descriptions in a natural way, we felt that these rather formal descriptions had to be converted into real English text. For example, the property ⟨`writerOf:The Pit and the Pendulum`⟩ can be written as "the writer of The Pit and the Pendulum." We created a program that converted properties from DBpedia into English using predefined mappings from properties to strings. All descriptions were post-edited by a native English speaker with experience in linguistics to remove any redundancies in the descriptions and to improve the fluency of the generated descriptions. The English speaker was not involved in the research and had no awareness of its aims.

Our null hypothesis is that *"there are no differences between algorithms in terms of the numbers of correctly identified referents,"* and similarly, *"there are no differences between algorithms in terms of their naturalness and quality."* We expected the descriptions produced by Human to perform best, as their descriptions are likely to contain enough information to unambiguously identify the referent. The descriptions extracted from DBpedia (algorithm DBP) are likely to perform poorly in terms of identification, as they are often quite general (e.g., *"a famous English writer"*).

5.2. Materials for Evaluating the Model

The names of targets were selected from two websites with lists of names of famous people[7]. We selected 100 names that were not used in our pilot experiments. The evaluation therefore contained 100 names and 500 descriptions in total, given that each referent was described using 5 sources: Alg1, Alg2, IA, DBP and Human.

We used a repeated-measures Latin square design in which each participant viewed a number of descriptions generated by each algorithm (within-subject design). To avoid presenting participants with too many descriptions, the 100 names were randomly assigned to 4 groups of 25. Each group was arranged into a Latin square so that each participant judged 5 descriptions generated by each of the 5 sources (25 descriptions per participant). Furthermore, each description was viewed three times, each time by a different participant.

The order in which descriptions were viewed might bias the results (e.g., seeing a description of Albert Einstein might make it easier to guess Niels Bohr), therefore the order of the descriptions was randomized for every participant. Each description was viewed by three participants.

5.3. Participants and Procedure for Evaluating the Model

The evaluation was carried out in accordance with the recommendations of the University of Aberdeen Handbook For Research Governance and approved by the College of Physical Sciences Ethics Review Board. Participants were informed that their participation was completely voluntary and that they could withdraw from the survey at any time for any reason. Participants were informed that the information was used solely for research purposes. No personal information would be shared with any third party. Participants who agreed with the conditions could proceed with the experiment.

The evaluation involved 60 participants (37 male, 22 female and 1 unspecified). In terms of highest achieved education, 26 of the participants had high school, 26 participants had an undergraduate degree and 8 participants had a postgraduate degree. Participants took on average 28 min to complete the task.

[7] www.biographyonline.net/people/famous-100.html and www.whoismorefamous.com/?fulllist=1 last retrieved 21st August 2015.

Once again, participants were recruited using MTurk. The experiment was advertised only to US MTurk workers who had at least 85% success rate (at least 85% of tasks that a worker submitted in the past were deemed acceptable by the requester). The reason for advertising only to US workers was to maximize the overlap between the knowledge of the participants and the knowledge captured by DBpedia. The famous people employed as target referents are famous in western culture, particularly in the US. The selection criteria ensured that participants were from essentially the same population as the participants who created the corpus of Section 3.

Participants were directed to a website that provided instructions. Participants were asked to provide some demographic information (age group, interests, etc.). After submitting this information, participants were shown detailed instructions on how to fill in their answers. After clicking on a button the first description appeared on the screen (**Figure 5**).

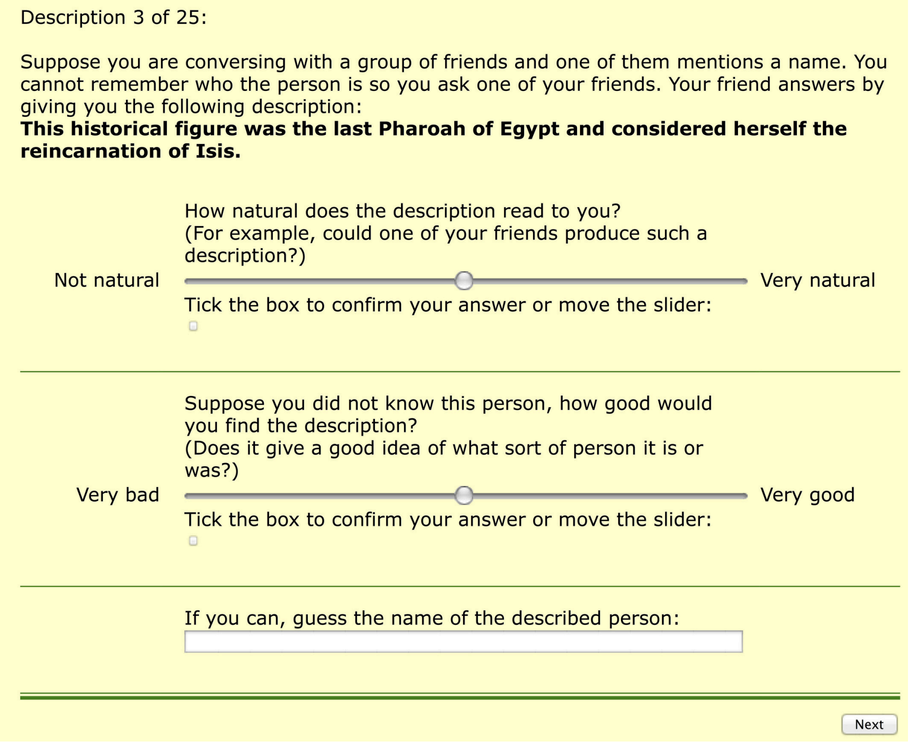
Participants had to judge the description and fill in the name of the referent if they could. The two judgment questions were: *"How natural does the description read to you? (For example, could one of your friends produce such a description?)"* and *"Suppose you did not know this person, how good would you find the description? (Does it give a good idea of what sort of person it is or was?)"*. These two judgments are referred to as *naturalness* and *quality* respectively. Participants provided ratings by moving sliders (similarly to Gatt et al., 2009). The sliders corresponding to each statement were set to the middle position and participants gave their judgment by moving each slider along the horizontal axis. The numerical values corresponding to the sliders were 1–100, but participants were not shown the number. If a participant wished to leave the slider in the middle (value 50), they had to tick the corresponding check box below the slider. This was done to prevent participants from accidentally leaving the slider in its original position without intending to offer a judgment.

Clicking the button "Next" sent participants to a page with the description and the name of the described person. Participants then had to choose their response from one of the options in **Figure 6**. The options were mutually exclusive and were used to gain more insight into the features of the descriptions that were produced:

Option 1 indicates successful identification. It was also used to filter out participants who did not take the task seriously. Any participant who provided a wrong name and selected this option was removed from the result set (as they were shown the correct name). The names provided by participants were checked against the actual referent names.

Option 2 was included for participants who experienced the "tip-of-the-tongue" event (ToT, Brown and McNeill, 1966). This option accounted only for about 3% of answers (Kutlak et al., 2013).

Option 3 accounts for situation where the algorithm selects information that is not known by listeners. If an algorithm generates descriptions that frequently lead to this situation, the



Description 3 of 25:

Suppose you are conversing with a group of friends and one of them mentions a name. You cannot remember who the person is so you ask one of your friends. Your friend answers by giving you the following description:
This historical figure was the last Pharoah of Egypt and considered herself the reincarnation of Isis.

How natural does the description read to you?
(For example, could one of your friends produce such a description?)

Not natural — Very natural

Tick the box to confirm your answer or move the slider:
□

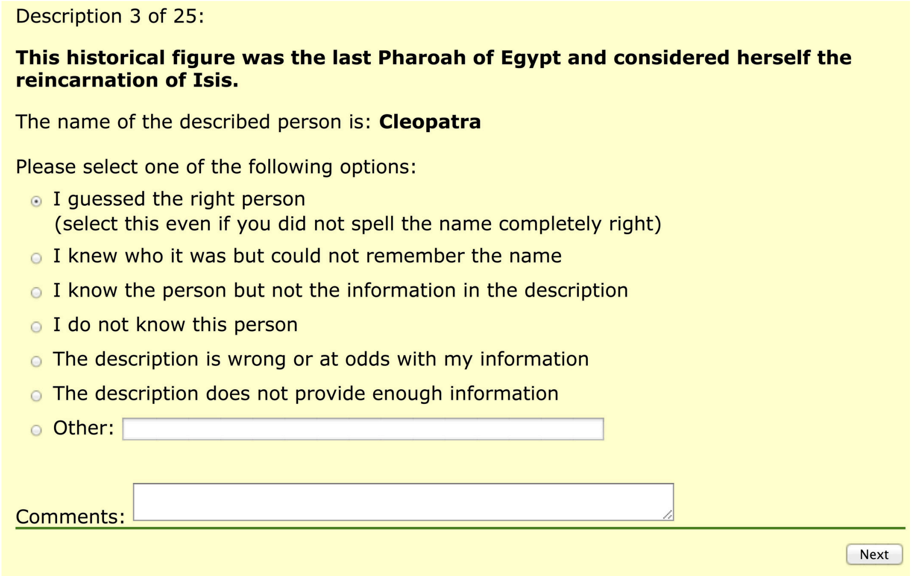
Suppose you did not know this person, how good would you find the description?
(Does it give a good idea of what sort of person it is or was?)

Very bad — Very good

Tick the box to confirm your answer or move the slider:
□

If you can, guess the name of the described person:

FIGURE 5 | Presentation of descriptions in the evaluation. Participants provided judgment of each description by moving the sliders. The box at the bottom of the page was used for providing the name.



Description 3 of 25:

This historical figure was the last Pharoah of Egypt and considered herself the reincarnation of Isis.

The name of the described person is: **Cleopatra**

Please select one of the following options:

- ● I guessed the right person (select this even if you did not spell the name completely right)
- ○ I knew who it was but could not remember the name
- ○ I know the person but not the information in the description
- ○ I do not know this person
- ○ The description is wrong or at odds with my information
- ○ The description does not provide enough information
- ○ Other:

Comments:

FIGURE 6 | Options shown to participants after guessing the name of the described person.

algorithm is probably selecting properties that are too difficult for people to recall (e.g., dates or numbers).

Option 4 was added to avoid lowering the score of good descriptions where participants do not know the target.

Option 5 covers situations in which some of the properties in the description are not true of the target (i.e., a participant believes to have knowledge that contradicts the information in the description).

Option 6 accounts for situations where an algorithm selects too few properties or where the selected properties are too general. For example, describing a target as *this person is an actor* is unlikely to allow identification of the target.

The participants had the chance to provide any other reason for not being able to identify the target (option 7) as well as providing comments.

TABLE 6 | Counts of selected answers for individual algorithms in the final evaluation.

	DBP	IA	Alg1	Alg2	Human	Total
Correct identification	68	58	89	100	180	495
Tip-of-the-Tongue (ToT)	21	17	24	24	28	114
Unknown target	44	64	61	61	51	281
Unknown properties	55	136	106	94	33	424
Underspecified	104	21	7	10	2	144
At odds with my information	2	3	8	6	3	22
Other	6	1	5	4	2	18
Total	300	300	300	299	299	1498

DBP are descriptions from DBpedia, IA is the Incremental Algorithm, Alg1 is the new algorithm that always selects three properties, Alg2 is the new algorithm that uses document retrieval as a termination heuristic and Human are descriptions produced by a native English speaker.

6. RESULTS OF EVALUATING THE MODEL

Table 6 contains frequencies of selected answers for each algorithm. Two answers were not saved due to a technical error. The χ^2 test compares the observed frequencies with expected frequencies based on the totals in each row and column. In order to focus on the differences in numbers of correctly identified referents, participants' answers were collated into two categories: *correct* and *incorrect*. Responses where participants did not know the referent were removed from the analysis. Where participants selected the Tip of the Tongue (ToT) option, they failed to provide the name of the referent, yet they did have the right person in mind; because this makes it difficult to say whether these answers were correct, we excluded them from our analysis. **Table 7** shows the collapsed counts where Incorrect Identification is the sum of the figures in the categories Unknown Properties, Underspecified, At odds with my information, and Other.

As we can see from the table, the descriptions generated by an English speaker outperform every other algorithm. The effect of algorithms was tested using the χ^2 statistics with the numbers from **Table 7**. The null hypothesis was rejected as the test showed significant differences: $\chi^2_{(4)} = 176.8, p < 0.001$. *Post-hoc* pairwise comparison shows statistically significant differences between algorithms IA and Alg2 [$\chi^2_{(1)} = 18.3, p < 0.001$], IA and Alg1 [$\chi^2_{(1)} = 10.1, p < 0.005$] and Alg2 and Human [$\chi^2_{(1)} = 56.8, p < 0.001$].

Even though no description contained the full name of the corresponding referent, some descriptions still contained a "clue" in the form of a property containing a part of the referent's name.

For example, *John Lennon* was described as *"This person wrote I Know I Know and was the topic of the musical Lennon."* In most cases, clues were the names of relatives (e.g., "the spouse of Victoria Beckham," "relative of Earl Woods") or names of related entities ("a member of The Jackson 5," "the creator of The Cosby Show").

To show the differences between the algorithms more clearly, we removed descriptions of all referents that contained such clues. This was the case for 29 out of the 100 entities and a total of 346 name guesses. **Table 8** contains the counts of correctly and incorrectly identified referents on the resulting subset of name guesses. The numbers of correctly identified referents differed significantly between the algorithms tested $\chi^2_{(4)} = 119, p < 0.001$. *Post-hoc* pairwise comparison resulted in the homogeneous subsets in **Table 9**.

The results show a difference between algorithms Alg2 and Alg1, suggesting that the content-based termination heuristic might increase the chances of successful identification. Note, however, that the difference was not statistically significant and more investigation is required to investigate this issue.

Table 10 shows mean *naturalness* and *quality* for each algorithm. While the differences in *naturalness* are small, the algorithms seem to differ substantially in terms of *quality*. Because each description was viewed 3 times, the data were aggregated so that the *naturalness* and *quality* ratings for each description were created by taking the mean of the 3 ratings. We performed two one-way analyses of variance with ALGORITHM as the independent variable. The main effect of ALGORITHM was significant on *naturalness* $[F_{(4,\ 495)} = 4.576, p < 0.005]$ and on *quality* $[F_{(4,\ 495)} = 40.23, p < 0.001]$. **Tables 11**, **12** show homogeneous subsets for *quality* and *naturalness* calculated by *post-hoc* Tukey test. Algorithms that do not share a letter are

TABLE 7 | Counts of correctly and incorrectly identified referents.

	DBP	IA	Alg1	Alg2	Human	Total
Correct identification	68	58	89	100	180	495
Incorrect identification	167	161	126	114	40	608
Total	235	219	215	214	220	1103
Proportion correct	0.29	0.26	0.41	0.47	0.82	0.45
Proportion incorrect	0.71	0.74	0.59	0.53	0.18	0.55
Correct/incorrect	0.41	0.36	0.71	0.88	4.50	0.81

The table also shows the proportions as well as the ration of correctly and incorrectly identified referents.

TABLE 8 | Counts of correctly and incorrectly identified referents when descriptions that contained a clue as to the identity of the referent were removed.

	DBP	IA	Alg1	Alg2	Human	Total
Correct identification	44	41	49	63	123	320
Incorrect identification	112	110	97	86	32	437
Total	156	151	146	149	155	757
Proportion correct	0.28	0.27	0.34	0.42	0.79	0.42
Proportion incorrect	0.72	0.73	0.66	0.58	0.21	0.58
Correct/incorrect	0.39	0.37	0.51	0.73	3.84	0.73

The table also shows the proportions as well as the ratio of correctly and incorrectly identified referents.

TABLE 9 | Homogeneous subsets for counts of *correctly identified referents*.

Algorithm				Correct	Total
Human	A			123	155
Alg2		B		63	149
Alg1		B	C	49	146
DBP			C	44	156
IA			C	41	151

Algorithms that do not share a letter are significantly different with p < 0.05.

TABLE 10 | Mean ratings and standard deviations for *quality* and *naturalness* for each algorithm in the final evaluation.

	DBP	IA	Alg1	Alg2	Human
Mean quality	43.570	57.857	67.600	66.786	77.552
Quality SD	27.385	20.630	16.670	18.051	15.570
Mean naturalness	61.953	61.927	62.110	61.495	70.311
Naturalness SD	18.587	17.294	17.758	18.655	15.556

Quality *refers to the statement: "Suppose you did not know this person, how good would you find the description?" and* naturalness *refers to the statement: "How natural does the description read to you?"*

TABLE 11 | Homogeneous subsets for *quality* calculated using *post-hoc* Tukey test.

Algorithm					Mean quality	*SD*
Human	A				77.6	15.6
Alg2		B			66.8	18.1
Alg1		B			67.6	16.7
IA			C		57.9	20.6
DBP				D	43.6	27.4

Algorithms that do not share a letter are significantly different at p < 0.05.

TABLE 12 | Homogeneous subsets for *naturalness* calculated using a *post-hoc* Tukey test.

Algorithm			Mean naturalness	*SD*
Human	A		70.3	15.6
Alg2		B	61.5	18.7
Alg1		B	62.1	17.8
IA		B	61.9	17.3
DBP		B	62.0	18.6

Algorithms that do not share a letter are significantly different at p < 0.05 statistically different with $p < 0.05$. As we can see, the human-produced descriptions were rated highest. The analysis suggests that descriptions produced by the new algorithms have higher "*quality*" than the ones produced by the IA and DBpedia.

7. GENERAL DISCUSSION

Our computational model addresses what we believe to be an interesting variant of the much-studied problem of reference production. The model does a better job addressing its task—producing descriptions of famous people to an unknown audience—than the Incremental Algorithm, both in terms of the numbers of correctly identified referents and in terms of the perceived quality of the descriptions generated. The structure of the model differs sharply from earlier ones. This is not only true in comparison to the algorithms proposed in practical Computational Linguistics (of which Siddharthan et al., 2011, developed in a context of automatic text summarization, is a good example), but also in comparison to algorithms developed in the tradition of REG. To ensure that our contributions are understood properly, it is worth re-stating some features of our approach.

Although some previous models of referring were constructed for situations in which hearers know less than speakers (Garoufi and Koller, 2014b; Paraboni and van Deemter, 2014b), these models assume that *the speaker knows what the hearer knows*. In many situations, however, speakers do not possess this information, for example when a journalist writes a newspaper article or a scientist a journal paper. Our model targets situations of this kind, where the key to success lies in the model's ability to make an educated guess concerning the knowledge of the reader. Additionally, we argue that these situations require that a guess is made about what information is likely to be distinctive for the hearer, and how much information the reader is likely to require.

Our model differs from earlier REG models because all three heuristics composing our model make use of pre-existing open-source data, rather than information that is hand crafted by researchers interested in reference. We believe that this lends additional interest to our model, because hand-crafting might accidentally benefit some algorithms over others. The use of open-source data is now well established in Computational Linguistics, but it has not been applied to the generation of referring expressions before.

Note, furthermore, that some key features of existing REG algorithms do not feature in our model. For example, since termination cannot be based on the criterion most often employed in REG (namely, that all distractors have been removed), we have had to find a different approach to this problem (Section 4.3). Similar observations can be made about discriminatory power (DP), a concept that had to be combined with information retrieval techniques to make it applicable to a situation in which the set of distractors is not know, and modified in light of the frequencies found in our data.

Some difficult questions are worth raising briefly. First, does our model have psychological reality, or is it merely a *product model* in the sense of Section 2? On the one hand, it is clearly a product model, since our final evaluation (Section 5) looked only at the output of the model, disregarding the actual production process. On the other hand, our tests of the individual heuristics do suggest that human speakers would be able to carry out these tests. Consider the Knowledge Heuristic, for example. Speakers evidently do not use Google to perform the kind of tests of which this heuristic makes use. Yet it may not be entirely implausible that speakers encounter, over the course of their lives, a large amount of text that is in some ways similar to (a suitable part of) the world-wide web, consequently considering the web as a model of human knowledge might not be ridiculous. The Knowledge Heuristic is no perfect model of human speaking, but it may be our best tool for capturing one aspect of it. Similar things might be said about Unexpectedness and Termination.

Should the Knowledge Heuristic be seen as a model of Common Ground? The experiment in Section 4.1 did not look at deeper levels of epistemic embedding (as in the classic notion of Common Ground of Section 1): at most, this experiment established that the heuristic predicts (approximately) what hearers know. On the other hand, if it is true, as is generally assumed, that reference rests on Common Ground, then our final evaluation—which suggests that the descriptions generated are effective and at least somewhat natural—suggests that, despite its relative simplicity, the Knowledge Heuristic makes reasonable predictions concerning (communal) Common Ground itself.

An example may make this clearer. Suppose it is debatable who invented the printing press: most Americans believe that this was done by Mr. X, but most Chinese believe it was done by Mr. Y. An American speaker who addresses a Chinese audience might choose, politely, to refer to Mr. Y (i.e., the Chinese scientist) as "the inventor of the printing press." However, if her Chinese audience knows the speaker to be American, then they might misunderstand this description as referring to Mr. X (the American scientist), because they know that this is who the speaker believes the inventor of the printing press to be. Perhaps the success of our model can be seen as confirmation of Clark's idea that *communal* Common Ground tells us something about "real" (i.e., epistemically complex) Common Ground, and not just about a speaker's assessment of a hearer's knowledge.

Even though we have focussed on the production of referring expressions, it appears to us that elements of our proposal can be put to other uses. Consider Common Ground, for instance. To the extent that our Knowledge Heuristic is able to predict what facts a member of a particular community is likely to know, it is potentially relevant for many areas, such as journalism, advertising, and creative writing, because in all these areas it is important to assess what an unknown hearer is likely to know.

In a computational setting, the Knowledge heuristic is applicable to a key problem that arises in many Information Presentation systems, namely to decide what information should be provided to the user. In Natural Language Generation, this is known as the problem of general Content Selection (Reiter and Dale, 2000). This time the Knowledge Heuristic could be "used in reverse," selecting information that is *least* (rather than most) likely to be known, hence most worthy of being added to the reader's store of information (cf., Section 2.3, where Information Sharing is discussed). Similar observations can be made about the other two factors explored in our model. For example, many Content Selection approaches in Natural Language Generation aim to select interesting information; our Unexpectedness Heuristics could help.

Having said this, we acknowledge that we have merely put the first step on the road to understanding how the different factors may be estimated. For example, one could test the performance of our heuristics in specialist domains, for instance involving an audience of experts in some area of public life (say, football, or ballet), if a corpus of texts representing the knowledge of this audience can be found. Likewise, it would be interesting to investigate how the model fares at describing companies or geographic locations (rather than famous people).

A difficult question is how our approach might generalize to more complex types of information. At least in its implementation detail, the approach is difficult to extend to logically complex information: it is one thing to search a set of documents for the fact that Ernest Hemingway was American, for instance, (an atomic fact) than to search for the fact that he wrote *three novels*, or that he wrote *more novels than short stories*. This information may well be part of the Common Ground of a given community, but our computational model is not yet able to find it.

On a final, theoretical note, our Knowledge Heuristic can be seen as a first step toward a computational model of Herb Clark's Communal Common Ground. Our approach suggests, moreover, that it might be useful to extend the notion of Common Ground beyond its original conception, taking into account not only what speakers and hearers know, but also what they are interested in. After all, in communication, the interlocutors' interests are as important as their knowledge.

FUNDING

The research reported in this article is based on the Ph.D. project of Dr. RK, which was funded by the Scottish Informatics and Computer Science Alliance (SICSA). KvD acknowledges support from the EPSRC under the RefNet grant (EP/J019615/1).

REFERENCES

Appelt, D. (1985). Planning english referring expressions. *Artif. Intell.* 26, 1–33. doi: 10.1016/0004-3702(85)90011-6

Arts, A. (2004). *Overspecification in Instructive Texts*. PhD thesis, Tilburg University.

Atkinson, R., and Shiffrin, R. (1968). "Human memory: a proposed system and its control processes," in *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 2, eds K. Spence and J. Spence (New York, NY: Academic Press), 89–195.

Beaver, D. (1997). "Presupposition," in *Handbook of Logic and Language*, eds J. van Benthem and A. ter Meulen (Amsterdam: North Holland Publishing Co.), 939–1009. doi: 10.1016/B978-044481714-3/50022-9

Belke, E., and Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *Eur. J. Cogn. Psychol.* 14, 237–266. doi: 10.1080/09541440143000050

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol.* 22, 1482–1493. doi: 10.1037/0278-7393.22.6.1482

Bromme, R., Rambow, R., and Nückles, M. (2001). Expertise and estimating what other people know: the influence of professional experience and type of knowledge. *J. Exp. Psychol.* 7, 317–330. doi: 10.1037/1076-898x.7.4.317

Brown, R., and McNeill, D. N. (1966). The "tip of the tongue" phenomenon. *J. Verbal Learn. Verbal Behav.* 5, 325–337. doi: 10.1016/S0022-5371(66)80040-3

Brown-Schmidt, S. (2009). Partner–specific interpretation of maintained referential precedents during interactive dialog. *J. Mem. Lang.* 61, 171–190. doi: 10.1016/j.jml.2009.04.003

Clark, H. H., and Wilkes-Gibbs, D. (1986a). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Clark, H. H. (1996). *Using Language*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511620539

Clark, H. H., and Brennan, S. E. (1991). Grounding in communication. *Perspect. Soc. Shared Cogn.* 13, 127–149. doi: 10.1037/10096-006

Clark, H. H., and Marshall, C. (1981). "Definite reference and mutual knowledge," in *Elements of Discourse Understanding*, eds A. K. Joshi, B. L. Webber, and I. A. Sag (New York, NY: Cambridge University Press), 10–63.

Clark, H. H., and Murphy, G. L. (1982). "Audience design in meaning and reference," in *Language and Comprehension, Volume 9 of* Advances in Psychology, eds J.-F. L. Ny and W. Kintsch (Amsterdam: North-Holland Publishing Co.), 287–299.

Clark, H. H., and Wilkes-Gibbs, D. (1986b). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Dale, R. (1989). "Cooking up referring expressions," in *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics* (Morristown, NJ: Association for Computational Linguistics), 68–75. doi: 10.3115/981623.981632

Dale, R. (1992). *Generating Referring Expressions: Building Descriptions in a Domain of Objects and Processes*. Cambridge, MA: MIT Press.

Dale, R., and Haddock, N. (1991). "Generating referring expressions involving relations," in *Proceedings of the Fifth Conference on European Chapter of the Association for Computational Linguistics* (Morristown, NJ: Association for Computational Linguistics), 161–166. doi: 10.3115/977180.977208

Dale, R., and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902_3

Engelhardt, P. E., Bailey, K. G., and Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. New York, NY: Wiley.

Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336, 998. doi: 10.1126/science.1218633

Fussell, S. R., and Krauss, R. M. (1992). Coordination of knowledge in communication: effects of speakers' assumptions about what others know. *J. Pers. Soc. Psychol.* 62, 378–391. doi: 10.1037/0022-3514.62.3.378

Garoufi, K., and Koller, A. (2014a). Generation of effective referring expressions in situated context. *Lang. Cogn. Neurosci.* 29, 986–1001. doi: 10.1080/01690965.2013.847190

Garoufi, K., and Koller, A. (2014b). Generation of effective referring expressions in situated context. *Lang. Cogn. Process.* 29, 1–16. doi: 10.1080/01690965.2013.847190

Gatt, A., and Belz, A. (2010). "Introducing shared task evaluation to nlg: the TUNA shared task evaluation challenges," in *Empirical Methods in Natural Language Generation*, eds E. Krahmer and M. Theune (Berlin; Heidelberg: Springer), 264–293.

Gatt, A., Belz, A., and Kow, E. (2009). "The tuna-reg challenge 2009: overview and evaluation results," in *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation* (Morristown, NJ: Association for Computational Linguistics), 174–182. doi: 10.3115/1610195.1610224

Geng, L., and Hamilton, H. J. (2006). Interestingness measures for data mining: a survey. *ACM Comput. Surv.* 38:9. doi: 10.1145/1132960.1132963

Gupta, S., and Stent, A. (2005). "Automatic evaluation of referring expression generation using corpora," in *Proceedings of the Workshop on Using Corpora for Natural Language Generation* (Brighton: Citeseer), 1–6.

Heller, D., Skovbroten, K., and Tanenhaus, M. (2012). To name or to describe: shared knowledge affects referential form. *Top. Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01182.x

Hodges, J., Yie, S., Reighart, R., and Boggess, L. (1996). An automated system that assists in the generation of document indexes. *Nat. Lang. Eng.* 2, 137–160. doi: 10.1017/S1351324996001325

Holden, J., and van Orden, G. (2009). Dispersion of response times reveals cognitive dynamics. *Psychol. Rev.* 2, 318–342. doi: 10.1037/a0014849

Horacek, H. (2005). "Generating referential descriptions under conditions of uncertainty," in *10th European workshop on Natural Language Generation (ENLG-2005)* (Aberdeen), 58–67.

Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117. doi: 10.1016/0010-0277(96)81418-1

Isaacs, E. A., and Clark, H. H. (1987). References in conversation between experts and novices. *J. Exp. Psychol. Gen.* 116, 26–37. doi: 10.1037/0096-3445.116.1.26

Keller, F., and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Comput. Linguist.* 29, 459–484. doi: 10.1162/089120103322711604

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89, 25–41. doi: 10.1016/S0010-0277(03)00064-7

Kingsbury, D. (1968). *Manipulating the Amount of Information Obtained from a Person Giving Directions.* Unpublished Honors Thesis, Department of Social Relations, Harvard University.

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Koolen, R., Krahmer, E., and Theune, M. (2012). "Learning preferences for referring expression generation: effects of domain, language and algorithm," in *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference* (Utica, IL: Association for Computational Linguistics), 3–11.

Krauss, R. M., and Fussell, S. R. (1991). Perspective-taking in communication: representations of others' knowledge in reference. *Soc. Cogn.* 9, 2–24. doi: 10.1521/soco.1991.9.1.2

Kutlak, R., van Deemter, K., and Mellish, C. (2011). "Audience design in the generation of references to famous people," in *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (Boston, MA).

Kutlak, R., van Deemter, K., and Mellish, C. S. (2013). "Generation of referring expressions in large domains," in *Proceedings of the Workshop on Production of Referring Expressions: Bridging the Gap between Empirical, Computational and Psycholinguistic Approaches to Reference (PRE-CogSci'13)* (Berlin).

Lane, L. W., Groisman, M., and Ferreira, V. S. (2006). Don't talk about pink elephants! : speakers' control over leaking private information during language production. *Psychol. Sci.* 17, 273–277. doi: 10.1111/j.1467-9280.2006.01697.x

Levelt, W. (1989). *Speaking : From intention to articulation.* A Bradford book; ACL-MIT Press series in natural-language processing. London: MIT Press.

Lewis, D. (1979). Scorekeeping in a language game. *J. Philos. Logic* 8, 339–359. doi: 10.1007/BF00258436

Nickerson, R. S., Baddeley, A., and Freeman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychol.* 64, 245–259. doi: 10.1016/0001-6918(87)90010-2

Paraboni, I., and van Deemter, K. (2014a). Reference and the facilitation of search in spatial domains. *Lang. Cogn. Neurosci.* 29, 1002–1017. doi: 10.1080/01690965.2013.805796

Paraboni, I., and van Deemter, K. (2014b). Reference and the facilitation of search in spatial domains. *Lang. Cogn. Neurosci.* 29, 1002–1017. doi: 10.1080/01690965.2013.805796

Passonneau, R. J. (1996). Using centering to relax gricean informational constraints on discourse anaphoric noun phrases. *Lang. Speech* 39, 229–264.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 89–110. doi: 10.1515/ling.1989.27.1.89

Reiter, E., and Dale, R. (2000). *Building Natural Language Generation Systems*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511519857

Rosch, E. (1978). "Principles of categorization," in *Cognition and Categorization*, eds E. Rosch and B. B. Lloyd (Hillsdale, NJ: Lawrence Erlbaum), 27–48.

Siddharthan, A., and Copestake, A. (2004). "Generating referring expressions in open domains," in *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (Morristown, NJ: Association for Computational Linguistics), 407. doi: 10.3115/1218955.1219007

Siddharthan, A., Nenkova, A., and McKeown, K. (2011). Information status distinctions and referring expressions: an empirical study of references to people in news summaries. *Comput. Linguist.* 37, 811–842. doi: 10.1162/coli_a_00077

Stalnaker, R. (1978). "Pragmatic presuppositions," in *Semantics and Philosophy*, eds M. Munitz and P. Unger (New York, NY: New York University Press), 197–213.

Stubbs, J. B., and Tucker, G. R. (1974). The cloze test as a measure of english proficiency. *Modern Lang. J.* 58, 239–241. doi: 10.1111/j.1540-4781.1974.tb05105.x

Sun, R. (2008). *The Cambridge Handbook of Computational Psychology*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511816772

Turney, P. (2001). "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," in *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)* (Freiburg). doi: 10.1007/3-540-44795-4_42

van Deemter, K. (2016). *Computational Models of Referring: A Study in Cognitive Science*. Cambridge, MA; London: MIT Press.

van Deemter, K., Gatt, A., van der Sluis, I., and Power, R. (2012). Generation of referring expressions: assessing the incremental algorithm. *Cogn. Sci.* 36, 799–836. doi: 10.1111/j.1551-6709.2011.01205.x

van Gompel, R. P., Gatt, A., Krahmer, E., and van Deemter, K. (2012). "PRO: a computational model of referential overspecification," in *Proceedings of the Architectures and Mechanisms for Language Processing (AMLaP) Conference [abstract]* (Riva del Garda).

Vanderschraaf, P., and Sillari, G. (2009). "Common knowledge," in *The Stanford Encyclopedia of Philosophy*, ed E. N. Zalta, Spring 2009 Edn. Available online at: http://plato.stanford.edu/archives/spr2009/entries/common-knowledge/

Viethen, J., and Dale, R. (2010). "Speaker-dependent variation in content selection for referring expression generation," in *Proceedings of the 8th Australasian Language Technology Workshop* (Melbourne), 81–89.

Winograd, T. (1972). Understanding natural language. *Cogn. Psychol.* 3, 1–191. doi: 10.1016/0010-0285(72)90002-3

Wu, S., and Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cogn. Sci.* 31, 169–181. doi: 10.1080/03640210709336989

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 17 December 2015
doi: 10.3389/fpsyg.2015.01888

Global Similarities and Multifaceted Differences in the Production of Partner-Specific Referential Pacts by Adults with Autism Spectrum Disorders

Aparna Nadig[1,2]*, *Shivani Seth*[1] *and Michelle Sasson*[1]

[1] *School of Communication Sciences and Disorders, McGill University, Montreal, QC, Canada,* [2] *Centre for Research on Brain, Language and Music, McGill University, Montreal, QC, Canada*

Over repeated reference conversational partners tend to converge on preferred terms or *referential pacts*. Autism spectrum disorders (ASD) are characterized by pragmatic difficulties that are best captured by less structured tasks. To this end we tested adults with ASD who did not have language or intellectual impairments, and neurotypical comparison participants in a referential communication task. Participants were directors, describing unlexicalized, complex novel stimuli over repeated rounds of interaction. Group comparisons with respect to *referential efficiency* showed that directors with ASD demonstrated typical lexical entrainment: they became faster over repeated rounds and used shortened referential forms. ASD and neurotypical groups did not differ with respect to the number of descriptors they provided or the number of exchanges needed for matchers to identify figures. Despite these similarities the ASD group was slightly slower overall. We examined *partner-specific effects* by manipulating the common ground shared with the matcher. As expected, neurotypical directors maintained referential precedents when speaking to the *same* matcher but not with a *new* matcher. Directors with ASD were qualitatively similar but displayed a less pronounced distinction between matchers. However, significant differences and different patterns of reference emerged over time; neurotypical directors incorporated the *new* matcher's contributions into descriptions, whereas directors with ASD were less likely to do so.

Keywords: autism spectrum disorders, lexical entrainment, referential precedent, referential pact, partner-specificity, common ground, audience design, language production

Edited by:
Kees Van Deemter,
University of Aberdeen, UK

Reviewed by:
Marc Swerts,
Tilburg University, Netherlands
William S. Horton,
Northwestern University, USA

***Correspondence:**
Aparna Nadig
aparna.nadig@mcgill.ca

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *08 August 2015*
Accepted: *23 November 2015*
Published: *17 December 2015*

Citation:
Nadig A, Seth S and Sasson M (2015) Global Similarities and Multifaceted Differences in the Production of Partner-Specific Referential Pacts by Adults with Autism Spectrum Disorders. Front. Psychol. 6:1888. doi: 10.3389/fpsyg.2015.01888

INTRODUCTION

Autism Spectrum Disorders (ASD) are a group of neurodevelopmental disorders or conditions currently defined in the DSM-V by impairments in social communication and interaction alongside the presence of restricted and repetitive patterns of behaviors and interests (American Psychiatric Association, 2013). Perhaps the most noticeable communication difficulties people with ASD experience revolve around initiating and maintaining reciprocal conversation, which requires a host of pragmatic skills (Volden and Lord, 1991; Capps et al., 1998; de Villiers et al., 2007; Nadig et al., 2010). In the 1980s Baron-Cohen and colleagues proposed a compelling explanation for these difficulties, centering on *impaired theory of mind* or the ability to understand other's mental states and understand that these can differ from one's own (Baron-Cohen et al., 1985; Baron-Cohen, 1989). This account was based on early reports of impaired or even "absent" theory of mind in children with ASD and continues to be highly influential, though primarily outside the field of autism (cf. Klin et al., 1992; Frith and Happe, 1994; Mottron et al., 2006, for discussion of the limitations of theory of mind as a comprehensive account of autism).

In psycholinguistics the *impaired theory of mind* account sparked considerable interest in examining pragmatic abilities in ASD, often as a test case for models of reference that are rooted in considerations of *common ground* or the information mutually known to interlocutors (e.g., Clark and Marshall, 1981; Clark and Murphy, 1982; Clark, 1992). In particular, it is often hypothesized that if an aspect of language use is thought to rely on referencing another's mental state (e.g., Does my conversational partner know I call my computer "Titan" or do I need to refer to it as "my computer"?), then people with ASD should not be able to do it in a typical manner, or if they can, then this aspect of language use does not rely on theory of mind. We hope to demonstrate why this all or none approach is overly simplistic and unsubstantiated by current empirical evidence. Consequently a more nuanced view of the use of common ground in ASD is needed to inform models of reference, just as a multifaceted view has evolved on the use (Brown-Schmidt and Hanna, 2011) and representation (Brown-Schmidt, 2012) of common ground in the neurotypical population.

Although the impaired theory of mind account and conventional expectations hold that people with ASD should categorically lack sensitivity to common ground, research exploring whether people with ASD are sensitive to their conversational partner's perspective paints a more complex and gradient picture. Nadig et al. (2009) found that half of the children with ASD they tested showed reduced sensitivity to a partner's visual perspective when producing descriptions for objects that were either shared in common ground or visible only to them, as is commonly expected. However, the other half of children with ASD, who had higher formal language abilities, were indistinguishable from their typical peers with respect to reliance on common ground in this structured task. In a narrative task, DE Marchena and Eigsti (2016) manipulated common ground by having the listener share prior exposure (or not) to cartoon clips that were later narrated by participants. Their sample of adolescents with ASD showed sensitivity to common ground, communicating differently in its presence on a number of measures (explicit references of common ground, disfluencies, and independent ratings of communicative quality). Yet, while typically-developing adolescents showed a standard referential shortening effect, producing fewer words in narratives for listeners who shared exposure relative to those who did not, adolescents with ASD did not show this effect as a group. However, older ASD participants and those with better social skills performed similarly to their typical peers on this more open ended task. Taken together these findings demonstrate that reliance on common ground by children and adolescents with ASD is best viewed as delayed rather than absent, and that there is significant variation among people with ASD. Many speakers with ASD (who do not have language or intellectual impairment, as in these studies) are aware of differences in their partner's perspective, but are less adroit in addressing discrepancies in common ground in their spontaneous language use.

To date, one study has examined the negotiation of discrepant common ground in adults with ASD. Begeer et al. (2010) examined sensitivity to a partner's visual perspective during the comprehension of referential descriptions (employing a task similar to that used by Nadig et al., 2009) and found nearly identical performance between adults with and without ASD. Given the findings from children and adolescents reported above this is not surprising, as sensitivity to common ground increases with age and/or formal language abilities in ASD. Finally, Slocombe et al. (2013) used interactive tasks to investigate the alignment of lexical choice, spatial frame, and syntactic structure in adults with ASD without directly manipulating common ground information. They hypothesized lexical choice would involve *audience design* (Clark and Marshall, 1981) or the tailoring of language to the knowledge or competence level of a conversational partner to promote successful communication (e.g., Bortfeld and Brennan, 1997; Branigan et al., 2011). To examine lexical choice, Slocombe and colleagues used a referential communication task where a confederate described familiar pictures using rare names (e.g., chapel rather than church), and then measured whether participants would "align" with this uncommon name when later referring to the same picture. Contrary to the authors' predictions, they found that adults with ASD (specifically Asperger's Disorder using DSM-IV criteria) were as likely as comparison participants to use the uncommon name. They also aligned with their conversational partners with respect to syntactic structure and spatial frame of reference (Slocombe et al., 2013). In interpreting these findings, both Begeer et al. (2010) and Slocombe et al. (2013) highlight that the lack of group differences in their studies was likely due to the nature of the tasks employed, which were highly structured and goal-directed. There are a number of reasons why performance would be enhanced in structured language tasks vs. communication in real life. For one, the interaction is more predictable and the problem space is limited, so it may become easier to incorporate contextual information including a partner's perspective (Nadig et al., 2009). Begeer and colleagues proposed that over arousal and a focus on local rather than global processing that is observed in many individuals with ASD could be "neutralized in structured social interaction" (Begeer et al., 2010, p. 115). Importantly these authors (Nadig et al., 2009; Begeer et al., 2010; Slocombe et al., 2013) emphasize that audience-design effects from structured tasks are no less valid as evidence of reliance common ground during communication, but rather that structured tasks alleviate other factors and task demands that may normally interfere with the effective use of common ground in people with ASD. Nevertheless, these findings suggest that more open-ended tasks are required to capture difficulties with audience design that are commonly encountered by adults with ASD in daily life.

The goal of the current study was two-fold. First, we examined referential efficiency in the evolution of referential descriptions over multiple rounds of a communication game where directors, adults with or without ASD matched on verbal IQ, refer to novel tangram images (geometric shapes that are difficult to describe) so that matchers can identify them from an array of other tangrams (e.g., Krauss and Weinheimer, 1964, 1966; Clark and Wilkes-Gibbs, 1986; Krauss, 1987; Schober and Clark, 1989). We know from prior work using similar methods that referential efficiency improves over time through a process of *lexical entrainment* (Garrod and Anderson, 1987) where a preferred lexical form or *referential pact* (Brennan and Clark, 1996) is collaboratively agreed upon through proposals from the director as well as back channel responses from the matcher. To our knowledge the collaborative construction of novel referential terms has not previously been investigated in ASD. Though such a communication game is structured by definition, we view ours as a more open-ended task than those used previously due to a combination of factors: we examine the participant's open-ended production rather than comprehension of scripted instructions, stimuli is novel and unlexicalized, consequently it is difficult to describe and there is no closed set of options to choose from (e.g., common vs. rare), we investigate the evolution of descriptions over repeated rounds of reference, and finally we include an experimental manipulation (described below) to assess the impact of a change in partner, disturbing the structure that had been established. To examine participants' ability to entrain over time we analyze the duration of repeated rounds of the game with the same set of stimuli. We also investigate the number of descriptors (defined in Section Data Coding) directors produce when describing tangram stimuli, as well as the number of exchanges between director and matcher until the matcher is able to identify the target referent. Speakers with ASD are known to have difficulty providing the appropriate level of information for a given communicative situation, being over- or under-informative in different circumstances (cf. Volden et al., 1997; Dahlgren and Sandberg, 2008; Nadig et al., 2009), and to be stereotyped in their language use (Philofsky et al., 2007), which may make them less efficient in this collaborative task. However, previous data from adults with ASD without intellectual or language impairment, similar to our sample, demonstrates that lexical alignment is intact in this group (Slocombe et al., 2013). Therefore, we predicted few if any differences on measures of referential efficiency.

Second, we investigated the partner-specificity of any *referential pacts* established by manipulating the experience and thus the common ground shared with the matcher. Critically, in interactive settings (cf. Brown-Schmidt, 2009) lexical entrainment has been shown to be partner-specific. After conversational partners develop a referential pact, if a new partner who was not involved in entrainment is introduced, the entrained term is less likely to be maintained by directors (Brennan and Clark, 1996) or expected by matchers (Metzing and Brennan, 2003; Brown-Schmidt, 2009). Recent work with typically-developing children shows that children as young as 4 years old maintain referential precedents with their peers in a partner–specific manner (Köymen et al., 2014) and that 3- and 4-year-olds expect adult speakers to maintain referential precedents in a partner–specific manner (Matthews et al., 2010; Graham et al., 2014). The mechanisms underlying the comprehension of referential precedents is an area of active debate, at the heart of which is whether high-level common ground inferences or low-level memory mechanisms (episodic priming and encoding cues) best explain the effects (Brennan and Hanna, 2009; Brown-Schmidt, 2009; Shintel and Keysar, 2009; Kronmüller and Barr, 2015). The task we use stands somewhat outside this debate as it is a production task, and prior work on production used a different paradigm with familiar objects with known names rather than tangram stimuli (Brennan and Clark, 1996; Köymen et al., 2014). We view our task, where participants as directors need to create agreed upon terms for complex novel stimuli through interaction with a matcher, as one that inherently requires collaboration. Therefore, if partner-specific effects are found, they are likely to follow from considerations of whether a referential precedent is shared in common ground with a specific matcher or not, a point that will be returned to in the discussion.

We analyzed partner-specific effects by comparing expected differences across conditions in the duration of Round 1 vs. Round 4, where the new matcher was introduced in the *new* condition but the same matcher continued the game in the *same* condition. For a more precise measure of how directors may adapt descriptions to a *new* matcher, we examined the maintenance of the referential precedent from the prior round on critical Round 4 in the *same* vs. *new* conditions, as well as how they continued to interact with the matcher on Round 5, the end of the game with a given set of cards. Finally we explored whether these two variables were related on critical Round 4: Is the duration difference, which was expected to be a delay in the presence of a *new* matcher, related to whether directors continued to maintain the referential precedent or not? We predicted that neurotypical directors would maintain pacts with the same matcher but elaborate on the referential precedent or chose a different term when speaking to a new matcher, consistent with prior findings. When it comes to the ASD group, a staunch impaired theory of mind account would predict that they would show no difference between same and new matcher conditions, continuing to use the same descriptions regardless of differences in the common ground shared with their listener. However, given the findings reviewed above showing basic sensitivity to discrepancies in common ground in ASD, we expect this group to show some sensitivity to the change in partner but in a less pronounced way than the neurotypical group.

Finally, to obtain a direct measure of the collaborative nature of lexical entrainment (Clark and Wilkes-Gibbs, 1986) we examined how likely directors were to incorporate the matcher's proposals (when provided) for how to describe the figure. We predicted that matchers would suggest more descriptors on early rounds of discussing a figure, before a referential pact was established. We expected that participants with ASD may be less likely to engage in this collaborative behavior.

MATERIALS AND METHODS

Participants

Thirteen adults with ASD and 13 neurotypical adults (NT; from the general population with no known developmental disorders) were included in the sample. An additional 4 ASD participants were tested, but no video record of their sessions was available for analysis due to experimenter error. An additional 18 NT participants were tested but only those who could be closely matched to each of the ASD participants are included here. Participants with ASD were participating in a larger transition support service for young adults with ASD and were recruited through advertisements posted at local autism organizations, college offices for students with disabilities, and social service providers. The NT comparison participated in a longer 2 h testing protocol including the referential communication task presented here. They were recruited either through a psychology department subject pool, receiving partial course credit for participation, or through word of mouth and advertisements in the community, receiving $10 for participation. This study received ethics approval from the University of McGill Faculty of Medicine Institutional Review Board. All participants gave written informed consent in accordance with the Declaration of Helsinki.

Participants ranged from 18 to 29 years old; age did not differ significantly different between groups [ASD: $M = 22$ years 2 months, $SD = 4$ years 2 months, NT: $M = 21$ years 2 months, $SD = 11$ months, $t_{(1,\ 24)} = 0.90$, $p = 0.38$, $r = 0.17$]. Gender proportion was also similar across groups (ASD: 7 males, 6 females, NT: 5 males, 8 females, $\chi^2 = 0.62$, $p = 0.43$, $\varphi = 0.15$). To ensure the groups had similar verbal abilities, allowing for comparison of pragmatic abilities specifically on the referential communication task, they were administered the verbal subtests (Vocabulary and Similarities) of the Wechsler Abbreviated Sale of Intelligence (WASI; Wechsler, 2008) to obtain a measure of verbal IQ. Groups did not differ significantly with respect to verbal IQ [ASD: $M = 113$, $SD = 10$, NT: $M = 115$, $SD = 9$, $t_{(1,\ 24)} = 0.53$, $p = 0.60$, $r = 0.10$]. All but two participants (one from each group) were native speakers of English. Those who were not native speakers had been using English their daily life for 10 years or more and had completed secondary or university education in English, moreover they scored in the average range or higher on an English test of verbal IQ.

Community diagnoses of ASD were confirmed in our study by administration of the ADOS-2 module 4 (Lord et al., 2012), using the revised algorithm for module 4 (Hus and Lord, 2014) or, when possible, parent report of early autism symptoms using the Social Communication Questionnaire (SCQ, Rutter et al., 2003). Nine of 13 ASD participants met ASD criteria (i.e., scores of 8 or higher, $M = 13$, range = 8–23) on the ADOS-2 based on current functioning and the remaining four participants met criteria for ASD based on their early development, as reported by their parent on the SCQ (i.e., scores of 15 or higher), but fell short of meeting ADOS-2 criteria based on current functioning (having ADOS-2 scores from 5 to 7). Prior to the lab visit, participants in the NT group completed a demographic questionnaire asking if they had ever been diagnosed with a developmental disorder, and whether they had a first or second degree relative with ASD; potential NT participants meeting either of these criteria were excluded. Of potential NT participants who completed the questionnaire, two were excluded from participation.

Materials

Eighteen tangram figures were printed in black ink on white cardstock. Two sets of nine stimuli were used, one set resembled animals (Set A), and the second resembled people (Set B), see **Figure 1**. Cards were laminated and two copies were made of each card to have identical sets for the director and matcher. Two easel boards with a 3 by 3 numbered grid marked on them were used to present the stimuli. Velcro in each square of the grid and on the back of each card allowed the cards to be attached and removed from the easel.

Procedure

We employed a collaborative referential communication game to assess production, incorporating elements from two previous lines of research: spontaneous lexical entrainment while describing complex novel stimuli (Krauss and Glucksberg, 1969; Clark and Wilkes-Gibbs, 1986), and manipulation of the *same* vs. *new* partner when studying the use of referential precedents in interactive tasks (Brennan and Clark, 1996, Experiment 3; Metzing and Brennan, 2003; Brown-Schmidt, 2009).

Participants always played the role of director, describing tangram stimuli to an experimental confederate who acted as the matcher. To measure partner specific effects two different matchers (original or *same* and *new*) were introduced in the *new* matcher condition, details provided below. An experimenter who conducted the longer testing session introduced the participant to the *same* matcher, who was presented as a lab member who had been called to help with this particular task. The experimenter explained the task to the director and the *same* matcher concurrently, assuming no familiarity with the task. Matchers were undergraduate or graduate student research assistants, or

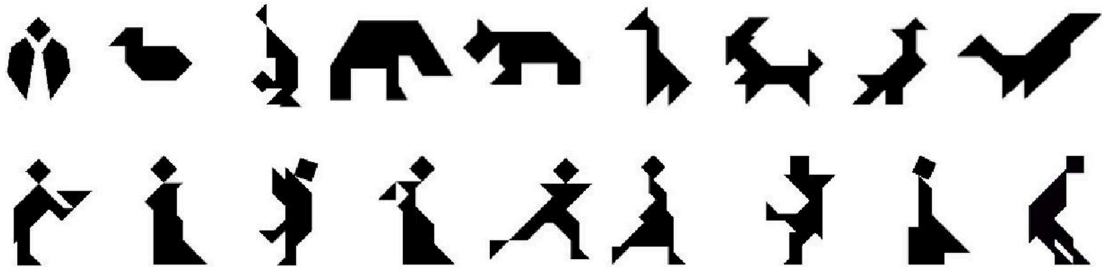


FIGURE 1 | Tangram stimuli used in our task, Set A (top row) and Set B (bottom row). Each set was used for one condition (*same vs. new* matcher).

in rare cases a faculty member if assistants were not available to fill all roles. For the sample reported here, in total 8 confederates played the role of the *same* matcher (median number of times playing this role = 4) and 10 confederates played the role of the *new* matcher (median number of times playing this role = 2.5). Matchers were instructed to respond naturally in the game and to ask for more information as required to complete the task. As the majority of matchers played the task only a few times over many months, and that the initial descriptions for each figure varied greatly across directors, the stimuli remained relatively new to them.

The director and matcher were seated across from each other at a table, each with a large easel in front of them so that neither could see the other person nor what was displayed on his/her easel. The experimenter (E), who conducted the longer session, introduced the task and how it was played, and was responsible for placing and removing stimuli cards and operating a videocamera. Before each round, E placed nine cards on the director's easel in a random order. E explained to both the director and matcher that she would put nine cards up on the easel, and that the matcher had an identical set of cards on the table in front of her. The director's task was to describe the cards in sequence (squares 1–9 were indicated on the board) so the matcher could place her cards to match the director's display. Three practice cards were used to familiarize dyads with the task. During the practice round it was reinforced that the director should move in order from square one to nine, that the matcher could ask questions at any time for clarification, and that the matcher should say "okay" or "got it" when she located the correct card as the director would not be able to see this.

Ten rounds of this game were played in total: five rounds with a given set of cards for each of the two experimental conditions (*same* matcher vs. *new* matcher). The director described nine cards, in sequence, on each round. At the end of the round, E shuffled the cards and placed them up on the director's easel in a random order, starting the next round.

The matcher's knowledge of referential precedents was manipulated as follows. In the *same* matcher condition, the director described the cards to the *same* matcher for three rounds. At this point the matcher said she forgot to tell her friend something next door and left. She returned after a minute, the game continued for rounds 4 and 5 with the *same* matcher.

In the *new* matcher condition the director also played the game with the *same* matcher for rounds 1–3. However, this time the *same* matcher said that she really needed to go to the bathroom and that her friend could step in for her. The *same* matcher left and the *new* matcher came in a minute later. E quickly introduced the *new* matcher to the game and its rules, and rounds 4 and 5 were played with the *new* matcher. Thus, the *new* matcher was also presented as a lab member, but one who was naïve to the game; the Experimenter introduced the game to this *new* matcher as if she had no prior knowledge of it.

A different set of cards (A or B in **Figure 1**) was used for each condition. Card set and order of condition were counterbalanced within each group by assigning each subsequent participant tested to one of four orders (e.g., Set A or Set B first, *same* or *new* condition first). Given an uneven sample size (13 in each group) and that more participants were tested than those included in the final sample, this resulted in card set and condition not being fully balanced. In the *same* matcher condition 8/13ASD participants and 7/13 NT participants received Set A, with the remaining receiving Set B; the opposite card set was used in the *new* matcher condition. For both ASD and NT groups 5/13 participants had the *same* matcher condition first, the remaining 8 had the *new* matcher condition first.

Data Coding

Data was transcribed from video recordings of the task. The *duration of each round* was recorded while transcribing. Tangram descriptions were divided into descriptors, defined as any noun or modifier describing the figure as a whole. Adjectives modifying a part of the figure were not counted as their own descriptor. For example, "the skater with his left leg stretched out and a diamond head" was coded as three *initial descriptors*: skater; left leg stretched out; diamond head. In this case, left was not considered a separate descriptor because it describes leg and not the skater. The *number of exchanges* or turns between director and matcher when working on a figure, from the initial description until the matcher located the card, were also coded. An exchange was defined as one description by the director, and one verbal response by the matcher. The response by the matcher could be a question or statement, or a confirmation ("Okay, got it" or "uh huh," participants were told during practice that they should confirm in this fashion since their partner could not see when he/she found the card).

Relation to the referential precedent of the prior round was coded into one of four categories. In doing so, determiners, prepositions and other function words were excluded when determining informational equivalence; thus "the lady who is walking" was considered equivalent to "walking lady" (Brennan et al., 2013). The *same* category was used when two descriptions were informationally equivalent. *Same-simplified* refers to descriptions that maintained the conceptualization of the previous round, but used fewer descriptors as is typical in lexical entrainment, apparently because fewer were needed when a referential pact had been established. For instance:

> Round 1: Director: The waiter with the triangle tray, facing right
> Round 2: D: The waiter with the tray
> Round 3: D: The waiter
> Round 4: D: Waiter
> Round 5: D: Waiter

In this example the description categories on Rounds 2 and 3 (with respect to the prior round) were *same-simplified* and description categories for Rounds 4 and 5 were same.

Sometimes directors would use the same conceptualization but add additional information or descriptors. This was coded as *same-expanded*. For example:

> Round 1: D: The sad dog
> Round 2: D: The sad dog, facing left

Occasionally the director would offer a conceptualizations that was completely different from the prior round, these were coded as *different.*

Round 1: D: The bird facing left with two triangle feet
Round 2: D: The giraffe

Incorporation of matcher's descriptors was coded as direct measure of the collaborative nature of lexical entrainment (Clark and Wilkes-Gibbs, 1986). For each of the 9 cards described and entrained upon in each condition (*same* or *new* matcher), how the director handled the matcher's spontaneous contributions regarding how to describe the figure were coded as follows: *yes* (matcher's descriptor incorporated by director on a subsequent round), *no* (no incorporation of descriptors suggested by matcher), or *N/A* (matcher did not propose any descriptors). This code was assigned at two time points for each card: for exchanges through the end of Round 3 (which always involved the *same* matcher in both conditions), and again for those from Round 4 through 5 (which involved a change in matcher in the *new* matcher condition). For instance, the following exchange in the *new* matcher condition received the code of *yes* for the Round 4 through 5 incorporation variable. It should be noted that this variable is likely affected by unmeasured differences with respect to the matcher's contributions (e.g., how plausible they were as descriptors, whether they offered a contrasting conceptualization or followed the director's conceptualization), since many confederates played the role of matcher and their only direction was to respond naturally to complete the task.

Participant 117
Round 4

Director: a four legged or two legged animal facing the right The head is a parallelogram and its back leg is a rectangle and the front legs look like paws
***New* Matcher:** does it look like an elephant if the parallelogram is a trunk?
Director: yeah, it does look like an elephant

Round 5

Director: an elephant facing right

Coding Reliability

A coding system was developed by the authors over multiple iterations of trying to capture the construct of referential precedent in the current production corpus involving the description of complex novel stimuli (as opposed to familiar basic and subordinate level terms, e.g., shoe and penny loafer, Brennan and Clark, 1996). The second author trained two additional undergraduate students, who were blind to the design and hypotheses of the study as well as group membership, on the final coding system via discussions and work on two training files until they reached consensus in their coding. Across variables, 20–33% of the participants in each group were double coded to calculate inter-coder reliability. For *number of initial descriptors*, correlations indicated very high agreement between both additional coders ($r = 0.96$ and 0.99) and the second author. *Number of exchanges* was calculated by an excel formula based on the cells where each director description and matcher response or question were entered in sequence. Reliability on referential *precedent categories* was measured by Cohen's unweighted kappa, which was reasonably high between each of the additional coders (Kappa = 0.90 and 0.78) and the second author. Finally, reliability for *Incorporation of matcher's descriptors* was obtained between the third author and an undergraduate student blind to study hypotheses and group membership. Cohen's unweighted kappa a was very high for incorporation by Round 3 (Kappa = 0.92) as well as by Round 5 (Kappa = 0.98).

RESULTS

Effect size is provided for each contrast using r for pairwise comparisons, for which a small effect is defined as 0.1, a medium effect is 0.3 and a large effect is 0.5, and with partial eta squared η_p^2 for ANOVA effects, which reflects the portion of unique variance on the dependent variable that is explained by the independent variable.

Referential Efficiency[1]

Round Duration

For a global analysis of whether lexical entrainment took place over the 5 rounds of each condition, we submitted data on round duration in seconds to a mixed ANOVA with round (1 through 5) and condition (*same* or *new* matcher) as within-subjects variables and group (ASD vs. NT) as a between subjects variable. As seen in **Figure 2**, there were strong main effects of round, $F_{(4,\ 96)} = 75.12$, $p < 0.001$, $\eta_p^2 = 0.76$, and of condition, $F_{(1,\ 24)} = 14.15$, $p = 0.001$, $\eta_p^2 = 0.37$. There was also a significant effect of group, $F_{(1,\ 24)} = 7.76$, $p = 0.01$, $\eta_p^2 = 0.24$. This was due to the ASD group having a higher average round duration (141 s) relative to the NT group (105 s) with a small effect size, $t_{(1,\ 234.12)} = 2.89$, $p < 0.001$, $r = 0.18$. There was a significant interaction between round and condition, $F_{(4,\ 96)} = 27.10$, $p < 0.001$, $\eta_p^2 = 0.53$. No other interactions were significant.

Remaining variables were averaged over the rounds of each condition and were analyzed using mixed repeated-measures ANOVAs with condition (*same* or *new* matcher) as a within-subjects factor and group (ASD, NT) as a between subjects factor.

Initial Number of Descriptors Per Figure

There was a significant main effect of condition (*same* vs. *new* matcher), $F_{(1,\ 24)} = 9.34$, $p = 0.005$, $\eta_p^2 = 0.28$. There was not a main effect of group, $F_{(1,\ 24)} = 0.49$, $p = 0.51$, $\eta_p^2 = 0.02$. Finally, there was no interaction between group and condition, $F_{(1,\ 24)} = 0.18$, $p = 0.70$, $\eta_p^2 = 0.01$. The NT group increased

[1] Variables presented here are considered to be constructs measuring referential efficiency (especially in the *same* matcher condition). However, given the partner manipulation in the experimental design, effects of condition (where differences are observed between *same* and *new* matcher) are best viewed as partner-specific effects that arose in Rounds 4 and 5 post partner switch.

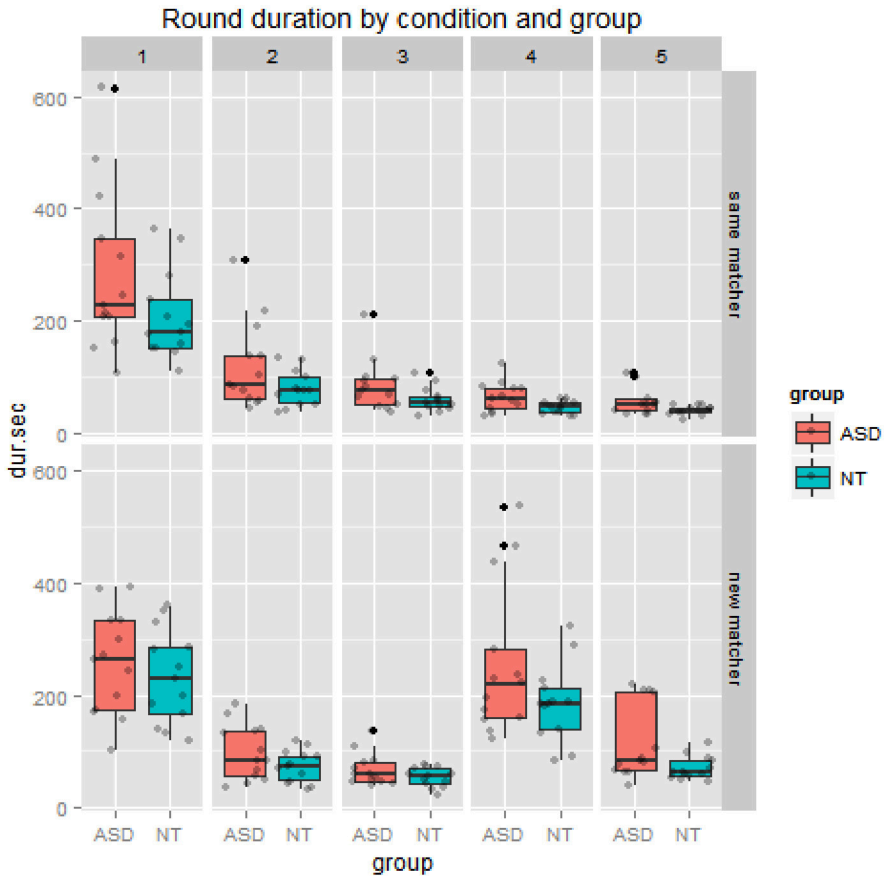


FIGURE 2 | The top panel shows the same matcher condition, where dyads in both groups get much faster over 5 rounds of discussing the figures, reflecting lexical entrainment. Results differ however in the *new* matcher condition in the bottom panel, where dyads in both groups show a disruption of lexical entrainment when a new matcher is introduced on the fourth round. Gray dots indicate jittered data points, black dots indicate outliers.

from a mean of 2.15 descriptors when speaking to the *same* matcher to 2.56 descriptors when speaking to the *new* matcher. The ASD group also increased across conditions, from a mean of 2.22 descriptors when speaking to the *same* matcher to 2.74 descriptors with the *new* matcher.

Number of Exchanges Per Round

There was a significant main effect of condition (*same* vs. *new* matcher), $F_{(1,\ 24)} = 11.13$, $p = 0.003$, $\eta_p^2 = 0.32$. Again there was no a main effect of group, $F_{(1,\ 24)} = 1.13$, $p = 0.30$, $\eta_p^2 = 0.05$. Finally, there was no interaction between group and condition, $F_{(1,\ 24)} = 0.60$, $p = 0.45$, $\eta_p^2 = 0.02$. **Figure 3** shows that the NT group increased from a mean of 1.32 turns when speaking to the *same* matcher to 1.53 turns when speaking to the *new* matcher. The ASD group also increased, from a mean of 1.45 turns with the *same* matcher to 1.59 turns with the *new* matcher.

Partner-Specific Adaptation

Round 4 was the critical point in the experiment; it was the first round after the original matcher left the room momentarily and returned in the *same* condition, or was replaced in the *new* matcher condition. We predicted that audience design effects would be seen most prominently at this point in the NT group. We predicted the ASD group would respond in a qualitatively similar way, showing some sensitivity to the change in matcher, but that they would smaller differences between the *same* and *new* matcher conditions than the neurotypical group.

Difference in Duration of Round 1 vs. 4

Through the process of lexical entrainment conversational partners typically speed up over repeated references to the same entity. To measure the extent to which the *new* matcher disrupted this process, controlling for baseline description speed, we calculated a difference score: duration of Round 1 minus the duration of Round 4, which is positive when dyads get faster over time. As would be expected from **Figure 2** above, there was a significant main effect of condition (*same* vs. *new* matcher), $F_{(1,\ 24)} = 38.48$, $p < 0.001$, $\eta_p^2 = 0.62$. As for other variables, there was no main effect of group $F_{(1,\ 24)} = 0.02$, $p = 0.88$, $\eta_p^2 = 0.00$. However, there was a marginally significant interaction between group and condition, $F_{(1,\ 24)} = 3.88$, $p = 0.06$, $\eta_p^2 = 0.14$. The NT group went from a mean decrease of 162 s when speaking to the *same* matcher to only 46 s when speaking to the

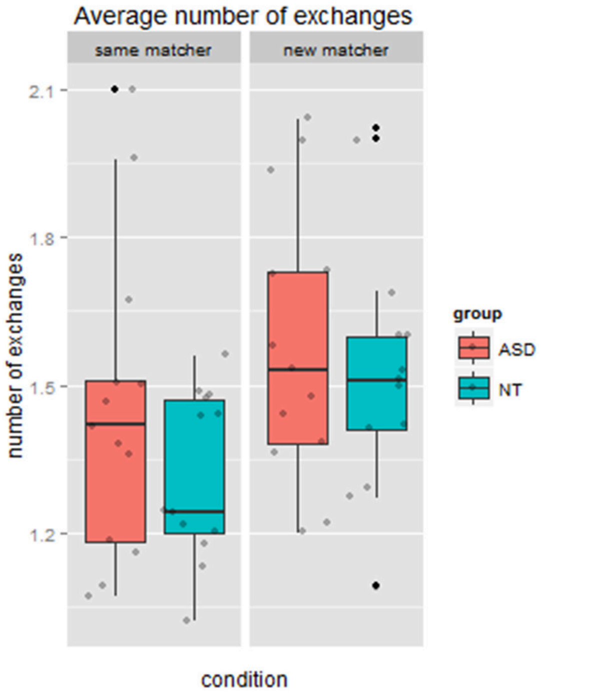


FIGURE 3 | Average number of exchanges required for the matcher to locate the figure, by condition and group.

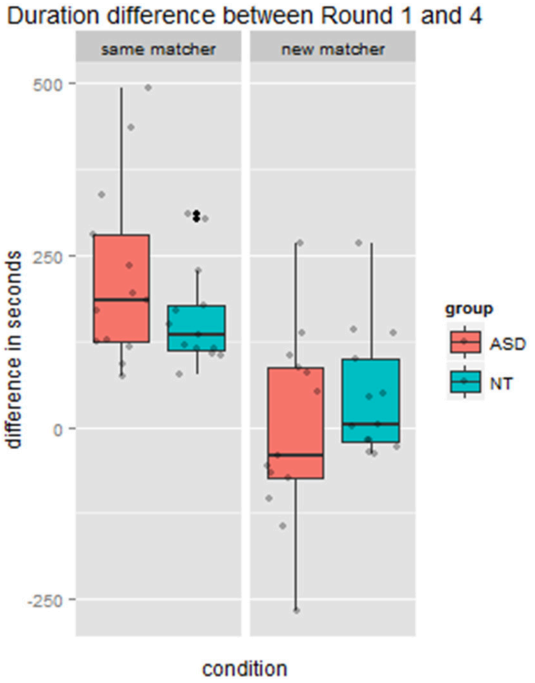


FIGURE 4 | Difference in duration of Round 1 vs. Round 4, where the new matcher switch occurred, by condition and group. Positive values indicate speeding up over four rounds of referring to the same figures, and a 0 duration difference indicates taking the same time on Round 4 as on the first presentation of the cards on Round 1.

new matcher. The ASD group went from a mean decrease of 221 s when speaking with the *same* matcher to an *increase* of 2 s with the *new* matcher. This pattern is depicted in **Figure 4**.

Maintenance of Referential Pact on Round 4

We predicted that on Round 4 directors in both groups would maintain the referential pact they had been using with the *same* matcher, that is, repeat the referential precedent or use a reduced form of it. For the *new* matcher we predicted that the NT group would spontaneously engage in audience design for the *new* matcher, who did not share knowledge of the referential precedent, by elaborating on it or using a different lexicalization. Finally, we predicted the ASD group would show less sensitivity to the *new* matcher by being more likely to maintain the referential pact they had established with the original matcher. **Figure 5** shows a complete tally of the types of descriptions given on critical Round 4 in relation to Round 3 descriptions.

Our analysis focused on maintenance of the referential precedent (including same-simplified and same descriptions). There was a significant main effect of condition (*same* vs. *new* matcher), $F_{(1,\ 24)} = 68.46$, $p < 0.001$, $\eta_p^2 = 0.74$. Once again there was no main effect of group, $F_{(1,\ 24)} = 1.61$, $p = 0.22$, $\eta_p^2 = 0.06$. There was, however, a marginally significant interaction between group and condition, $F_{(1,\ 24)} = 3.21$, $p = 0.08$, $\eta_p^2 = 0.12$. We followed this up with a planned comparison between groups in the *new* matcher condition specifically. In line with our prediction, the ASD group were marginally more likely to maintain the referential pact than the NT group, with a medium effect size, $t_{(1,\ 24)} = 2.03$, $p = 0.05$, $r = 0.37$. At the individual level, all 13 NT directors maintained 3 or fewer referential precedents on Round 4 with the *new* matcher, while 9/13 directors with ASD displayed the same pattern, but the remaining 4 maintained 4–8 referential precedents. As seen in **Figure 6** the NT group showed an extreme difference between conditions, maintaining referential precedents for a mean of 7 out of 9 figures when speaking to the *same* matcher, but only for 1.38 figures when speaking to a *new* matcher. The ASD group was less pronounced in this distinction, decreasing from a mean of 6.53 referential precedents maintained with the *same* matcher to 2.92 maintained with *new* matchers.

We also examined if Round 1 minus Round 4 duration difference was related to maintaining the referential pact on Round 4 in the *new* matcher condition specifically. We reasoned that that *maintaining the referential precedent* on this round may *slow the dyad's interaction*, since the *new* matcher lacked knowledge of the referential precedent. Consequently we expected that greater maintenance of referential precedents would be inversely related to duration difference (where positive values indicate speeding up). Results are shown in **Figure 7**. The correlation in the NT group was in the direction of our prediction but did not reach significance ($r = -0.20$, $p = 0.37$), likely because there was little variation in maintaining the precedent when speaking to the *new* matcher. There was a significant correlation in the ASD group ($r = 0.52$, $p = 0.02$), but in the opposite direction of our prediction. In fact, in cases where ASD participants maintained more precedents with the *new* matcher,

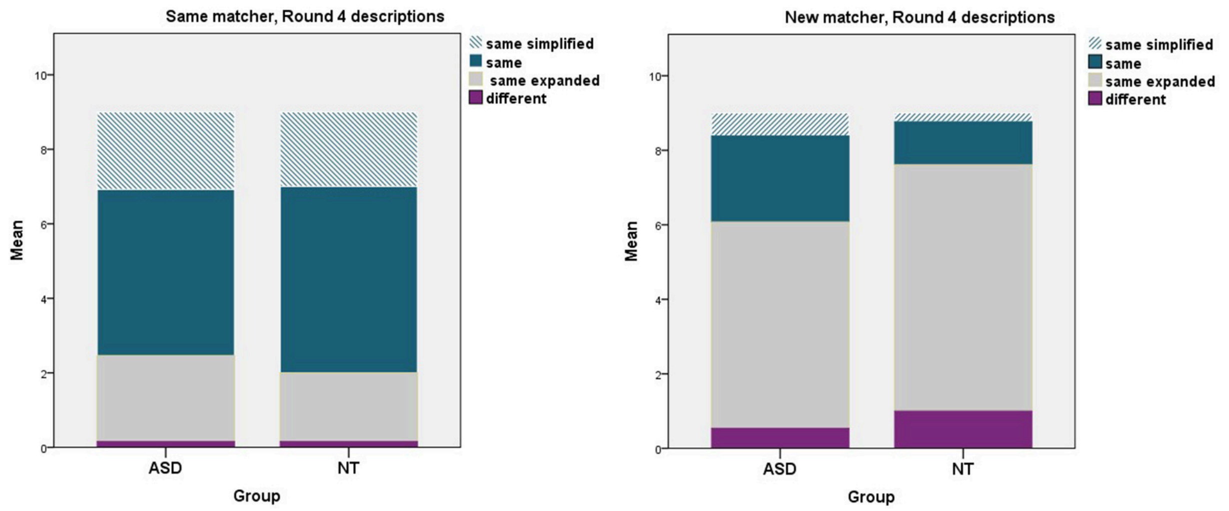


FIGURE 5 | Types of descriptions, with respect to the referential precedent of the prior round, given on critical Round 4 to the *same* matcher (left) vs. *new* matcher (right).

dyads got through Round 4 more quickly. Conversely, in cases where ASD directors tended not to maintain precedents with the *new* matcher, as NT directors did, dyads actually took longer to complete Round 4. Our interpretation of this finding is that it took ASD directors more time to adapt to the *new* matcher in the manner that NT directors did (by elaborating on the referential precedent or using a different term). It is also possible that the *new* matchers' responses contributed to this longer duration, however, since the *new matcher* had just started playing the game, and there were no group differences in referential efficiency measures on the part of the director, it is unlikely that matchers had a basis on which to respond differently to ASD vs. NT directors.

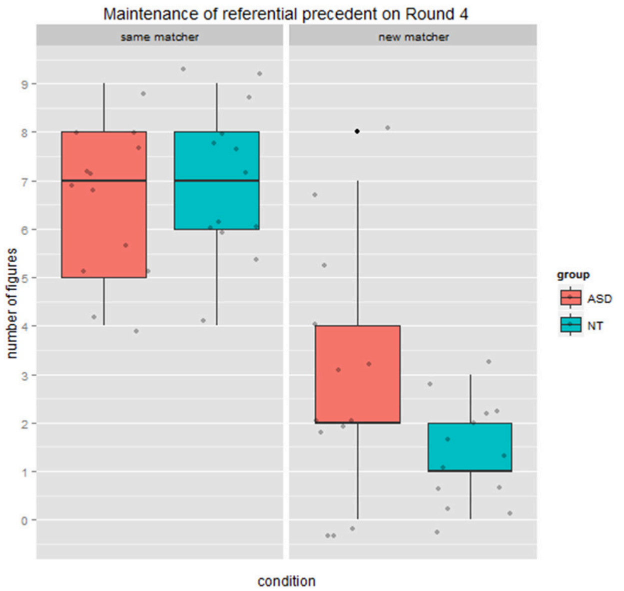


FIGURE 6 | Maintenance of referential precedent from Round 3 on Round 4, by condition and group.

Maintenance of Referential Pact on Round 5

To examine how entrainment would unfold in the presence of the *new* matcher we also examined maintenance of referential pacts from Round 4 on Round 5. There was a significant main effect of condition (*same* vs. *new* matcher) $F_{(1,\ 24)} = 27.13$, $p < 0.001$, $\eta_p^2 = 0.53$. Again there was no main effect of group, $F_{(1,\ 24)} = 0.14$, $p = 0.71$, $\eta_p^2 = 0.01$. Importantly, there was a significant interaction between group and condition, $F_{(1,\ 24)} = 4.98$, $p = 0.03$, $\eta_p^2 = 0.17$. **Figure 8** shows that the NT group maintained their referential precedent on the last round of the game for a mean of 7.92 of 9 figures with the *same* matcher, but for 5.23 figures when speaking to a *new* matcher. The ASD group was less divergent between conditions, decreasing from a mean of 7.30 referential precedents maintained with the *same* matcher to 6.23 maintained with *new* matchers.

Incorporation of Matcher Descriptors

This was a direct measure of how collaborative entrainment was, that is, whether directors incorporated descriptors suggested by the matcher on a subsequent round. Results are provided in **Table 1** below. The majority of data was missing for the *same* matcher condition, Round 4 through 5 because an entrained term was generally set and the *same* matcher tended not to suggest descriptors at this point, giving no opportunity for incorporation. Given this, analyses focused on the *new* matcher condition, which reflects partner-specific changes. A mixed ANOVA was conducted with subjects factor of time point (by end of Round 3, Round 4 through 5) and the between subjects factor of group. The effect of time point was not significant $F_{(1,\ 22)} = 1.83$, $p = 0.19$, $\eta_p^2 = 0.08$. There was however a significant main effect of group, $F_{(1,\ 22)} = 4.97$, $p = 0.04$, $\eta_p^2 = 0.18$. This was due

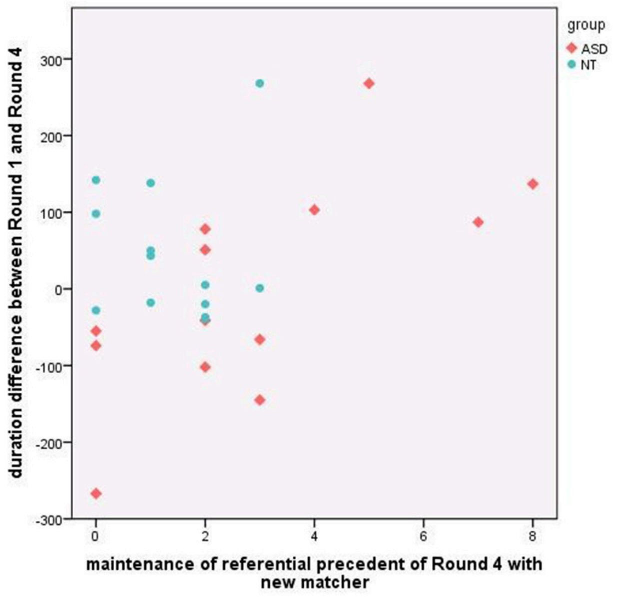


FIGURE 7 | Relation between maintenance of referential precedent on Round 4 with new matcher and the duration difference between Round 1 and Round 4 (positive values indicate speeding up over repeated reference).

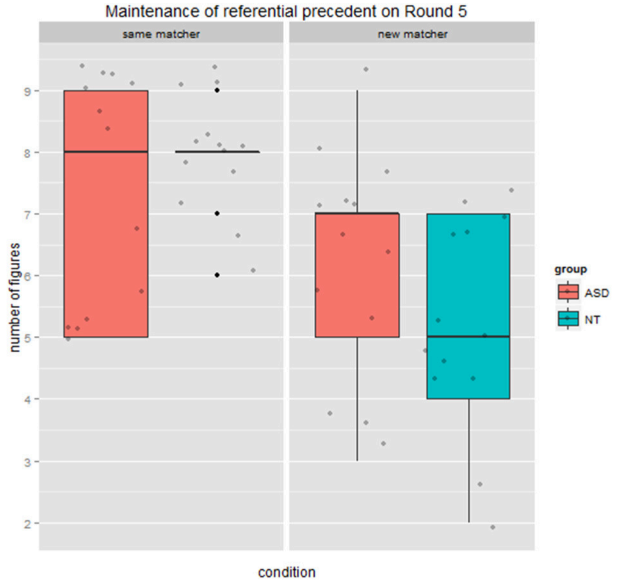


FIGURE 8 | Maintenance of referential precedent from Round 4 on Round 5, by condition and group. Note: The NT group displayed little variability, with an interquartile range of 1 that was too small to appear in this boxplot.

to the ASD group having a reduced tendency to incorporate the matcher's contributions (0.31) relative to the NT group (0.48). There was also a marginal interaction between group and time point, $F_{(1,22)} = 3.92$, $p = 0.06$, $\eta_p^2 = 0.15$. As seen in **Table 1**, this reflected the fact that, while the groups were similar in their incorporation of the matcher's descriptors when interacting with the original matcher until round 3, the ASD group became markedly less likely to do so than the NT group when interacting with the *new* matcher on rounds 4 through 5.

DISCUSSION

Our first set of findings on *referential efficiency* indicate that adults with ASD, who did not have language or intellectual impairment, were similar to a neurotypical comparison group with respect to the initial number of descriptors they used when describing tangram figures, and in the number of exchanges required for a matcher to find the figure they described. They also displayed the typical duration effect observed in lexical entrainment, becoming faster over time, to the same extent as the neurotypical group. These findings indicate that ASD group did entrain on lexical terms in this relatively open-ended task, rather than, for example, perseverating on the same description over five rounds. However, these similarities were observed in the presence of a global delay in completing the game: when directors were adults with ASD the game took significantly longer (on average 36 s longer per round) than when directors were neurotypical adults. This may have been due to differences in variables that we did not measure directly, for example the time taken to formulate descriptions of these complex novel figures, disfluencies when producing the descriptions, and/or the content of the description that may have led to the matcher to respond more slowly although there was no difference in the number of exchanges between director and matcher.

Our second set of findings focused on potential partner-specific effects in round duration and the maintenance of referential precedents. The pattern of results in the neurotypical group showed clear partner-specific effects, where round duration increased dramatically on Round 4 when the *new* matcher was introduced relative to when continuing with the *same* matcher. Interestingly, results from the ASD group belie a strong *impaired theory of mind* account which would predict no difference in round duration across matcher conditions. In fact, the ASD group showed the same condition effect as the neurotypical group, being delayed when the *new* matcher was introduced. Furthermore, there was a marginal interaction indicating that the ASD group had a tendency to be even more delayed by the change in matcher, rather than less delayed as we had expected. With respect to the maintenance of referential precedents, the neurotypical group exhibited robust partner-specific effects again, switching from maintaining the precedent almost all of the time with the *same* matcher to very rarely with the *new* matcher. The ASD group displayed a similar but less pronounced pattern, and were marginally more likely to continue to maintain precedents in critical Round 4 when interacting with a *new* matcher. We also found a counterintuitive correlation in the ASD group between maintenance of precedents on Round 4 when speaking to the new matcher and on duration difference: for those ASD participants who rarely maintained precedents (behaving like the neurotypical group), Round 4 trial duration was significantly longer, leading to negative difference scores. We suggest that this may reflect the effort required by directors with ASD

TABLE 1 | Incorporation of matcher's descriptors by condition, group and time point.

		Same matcher		New matcher	
		ASD	***NT***	***ASD***	***NT***
By end of round 3	Participants for whom the matcher suggested descriptors	100% (*n* = 13)	92.3% (*n* = 12)	100%(*n* = 13)	100% (*n* = 13)
	Proportion of the time directors incorporated matcher descriptors M *(SD)*	0.53 (0.38)	0.53 (0.32)	0.44 (0.31)	0.46 (0.25)
Round 4 through 5	Participants for whom the matcher suggested descriptors	38.5% (n = 5)	15.4% (n = 2)	84.6% (n = 11)	100% (n = 13)
	Proportion of the time directors incorporated matcher descriptors M *(SD)*	0.20 (0.18)	0.50 (0.70)	0.18 (0.19)	0.50 (0.27)

Red indicates cells where data is unreliable because the majority of data was missing.

to take the *new* matcher's common ground into account and formulate a more elaborated description as opposed to maintaining the referential precedent they had established with the original, *same* matcher. Importantly, these trends toward differences between groups on Round 4 were amplified on Round 5, where there was a significant interaction whereby the ASD group maintained referential precedents more often than the neurotypical (NT) group when interacting with the *new* matcher.

Data on the *Incorporation of matcher's descriptors* in the new matcher condition allows us to better understand the nature of this effect. Incorporation of descriptors offered by the matcher occurred close to half of the time *when working with the first matcher on Rounds 1 to 3,* among both NT participants and ASD participants. Clark and Wilkes-Gibbs (1986) proposed that when a director feels that the matcher is lacking information, he or she may choose to expand their description prior to being prompted to do so. Directors in our study, whether they were NT or had ASD, did so to a similar degree, incorporating new partner-specific descriptors given feedback that that the previous description was inadequate for the new matcher. This indicates another similarity in partner-specific effects.

However, a significant group difference emerged *when the last two rounds* of the *new* matcher condition were considered. This is the point where the first partner who was involved in lexical entrainment was replaced by a new partner who was viewing the tangrams for the first time. Analyses revealed that NT participants often gave elaborated descriptions on Round 4 when speaking to a *new* matcher (reflecting partner-specific adaptation). In addition, through Round 5, NT directors continued to incorporate terms suggested by the matcher half of the time on average, as they had on earlier rounds with the first matcher. In some cases they abandoned their initial formulation, producing a collaborative referential pact, as in this example, repeated from the methods section:

Participant 117
Round 4

> **Director:** a four legged or two legged animal facing the right The head is a parallelogram and its back leg is a rectangle and the front legs look like paws
> ***New* Matcher:** does it look like an elephant if the parallelogram is a trunk?
> **Director:** yeah, it does look like an elephant

Round 5

> **Director:** an elephant facing right

Consequently the director in this example did not maintain a referential precedent on Round 5. Sometimes, NT directors did not fully revise their description in favor of a distinct term offered by the matcher, but they added terms the matcher suggested into a collaborative referential pact and the Round 5 description was categorized as same-expanded, for instance:

Participant 124
Round 3

> (with ***Same*** **Matcher**)
> **Director:** lady sitting

Round 4

> **Director:** one where the lady is sitting
> ***New* Matcher:** sitting on a regular box?
> **Director**: yes

Round 5

> **Director:** the lady sitting on a box

In contrast, ASD participants were less likely to incorporate information offered by the *new* matcher (18% of the time) on Rounds 4 through 5. Instead, ASD participants tended to use the same descriptors they had on Round 4, or a simplification thereof. For example:

Participant 1333
Round 4

> **Director:** an arrow as a head
> ***New* Matcher:** is the top a square and then an upside down triangle?
> **Director:** yeah

Round 5

> **Director:** an arrow as a head
> ***New* Matcher:** and one foot is in the air?
> **Director**: yeah

Participant 1330
Round 4

> **Director:** reading

New **Matcher:** someone's reading?
Director: mmhmm
New **Matcher:** is he upside down?
Director: no it has a triangle on top
New **Matcher:** are there two shapes on either side that are the same? and each of the shapes have 5 sides?
Director: no someone reading facing to the left with a diamond

Round 5

Director: the one reading a book

These different patterns of incorporating *new* matcher perspectives on Rounds 4 through 5 in the case of NT directors, and a significantly decreased likelihood to do so by directors with ASD, likely gave rise to the significant interaction for maintenance of a referential pact on Round 5. Yet the groups did not differ with respect to incorporation of the *same* matcher's descriptors on earlier Rounds 1–3. Taken together, these findings demonstrate that adults with ASD do initially incorporate a partner's suggestions to the same degree as NT peers in the context of this task, which once again runs counter to an impaired theory of mind account. However, once a conceptualization and referential precedent has been established (in a collaborative manner), directors with ASD were less flexible in modifying the entrained upon term to accommodate the *new* matcher. Parallel findings were reported by Hala et al. (2007) who found that participants with ASD exhibited normal semantic priming of homographs in a first round of exposure, but not when a prime for the second meaning of the homograph was presented subsequently. Such findings can be explained by difficulties with inhibition or interference control in ASD (e.g., Geurts et al., 2014). Crucially, this is another example of communicative disruption in ASD, customarily attributed to theory of mind impairment, which actually follows from non-social difficulties (e.g., Nadig et al., 2010, where perseverative, self-contingent utterances in conversation were related to restricted and repetitive interest symptoms rather than social skills).

In summary, the adults with ASD in our study displayed largely typical effects of lexical entrainment in a collaborative game requiring them to develop referential descriptions for unlexicalized stimuli, but they took more time to do so than did neurotypical participants matched on verbal IQ. When their partner in the game changed to a *new* matcher, directors with ASD were sensitive to this change as a group, switching from maintaining referential precedents most of the time with the *same* matcher to significantly less often with the *new* matcher.

However, much more variability in making this switch was observed in the ASD than the neurotypical groups, leading to a marginal interaction with a medium effect size. Those directors with ASD who followed the neurotypical pattern of partner-specific adaptation in referential descriptions took significantly longer to complete the round with their partner, suggesting that this adaptation was effortful for them. This was coupled with a significant group difference in incorporating information proposed by the matcher, specifically by the *new* matcher at the end of the game. Neurotypical directors often added elements suggested by the *new* matcher, directors with ASD were resistant to do so at this point when a referential precedent was already established. Potentially due to this, on the last round of the game there was a significant interaction whereby directors with ASD maintained referential precedents with *new* matchers more often than did neurotypical directors. Collectively these represent a range of subtle but likely consequential differences in conversation that should be more pronounced in the context of real-life situations that are less predictable than this game.

We take these findings to indicate that adults with ASD have qualitatively typical patterns of basic lexical entrainment (as reported by Slocombe et al., 2013, for familiar, lexicalized stimuli), but take longer in this process. This global resemblance is offset by multifaceted differences in partner-specific aspects of reference, driven by a subgroup of participants with ASD. Interestingly, differences surfaced in two ways: a minority (4/13) of directors with ASD continued to maintain referential precedents when speaking to a *new* matcher who did not share history with it, while the majority of directors with ASD (9/13) made their descriptions more informative like neurotypical directors, but experienced delays in doing so. It seems that only the minority did not take notice of the *new* matcher's lack of common ground with the referential precedent, while the majority did notice but struggled to produce a description appropriate for their addressee.

We now return to the question of mechanisms that give rise to partner-specific effects in the use of referential pacts, and the debate between "cooperative" views where considerations of common ground guide language use (Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996; Metzing and Brennan, 2003; Brown-Schmidt, 2009) and low-level priming and encoding cue views where initial stages of language processing are independent of considerations of common ground (Horton and Gerrig, 2005; Kronmüller and Barr, 2007; Shintel and Keysar, 2007). Our study was not designed to address this debate, but we did find partner-specific effects on production at the first point possible (Round 4) after common ground was manipulated, consistently in neurotypical adults and in the majority of adults with ASD we tested. It is likely that both types of mechanisms are deployed in language use, simultaneously, in a graded fashion depending on a range of relevant factors including the strength of communicative goals, amount of cognitive load, and nature of stimuli (i.e., how entrenched linguistic forms are), (Brennan and Hanna, 2009; Branigan et al., 2011), in line with constraint-based models of language processing (MacDonald et al., 1994; Trueswell and Tanenhaus, 1994). The production paradigm we used pulls for cooperation, or at least mutually comprehensible reference, as there was no default way to describe the figures so descriptions were formulated over time through interaction (see excerpts above). Unlike comprehension studies using familiar, lexicalized stimuli where partner-specific effects can be explained by an expectation for speakers to be referentially consistent (e.g., Shintel and Keysar, 2007), it is difficult to imagine an explanation of our findings that does not entail considerations of the knowledge of the common ground available to the matcher.

A limitation of our study is its small sample size which resulted in some medium size effects not reaching significance. This is balanced by strengths in the rigorous matching of participants across groups and time intensive transcription and detailed coding required for task examining the evolution in production of lexical descriptions over time in a collaborative task. We investigated lexical entrainment between participants and experimental confederates given the logistical constraints of the partner manipulation, among others; an important direction for future work will be to examine lexical entrainment in ASD with naïve participants in both roles, as recommended by Kuhlen and Brennan (2013). Finally, we focused primarily on the director's contributions in this dyadic task; there remain many aspects of the matcher's influence on entrainment to be investigated. Matchers in this task were blind to the hypotheses of the study but not necessarily to group status, which may have become apparent through interaction, though it was not face to face. This gives rise to the possibility that confederate matchers may have engaged in audience design and communicative scaffolding, related to evaluations of the communicative competence (Bortfeld and Brennan, 1997; Branigan et al., 2011) of directors with ASD, which in turn contributed to some of the similarities observed between groups. Though possible we find this unlikely as the confederates were matchers in this game, as opposed to directors who took the lead in formulating descriptions, and the matcher had genuine informational needs as the terms used to describe the complex novel figures were highly idiosyncratic.

Crucially, this first investigation of partner-specific referential pacts in ASD resulted in a complex pattern of results that does not support a categorically *impaired theory of mind* account. The current findings reflect the communicative capacities of adults with ASD who do not have language or intellectual impairment; more pronounced group differences would be expected in children and more representative samples including individuals with ASD who have language and intellectual delays. Future work should explore the nature of two different patterns of partner-specific effects observed here in adults with ASD: not modifying descriptions in the presence of a *new* matcher who did not share common ground in a minority of participants, and adapting descriptions but this entailing a delay in the majority of participants. Further investigation is needed to examine the impact these relatively subtle differences on communication in the lives of people with ASD, including how they are perceived by various conversational partners.

FUNDING

This study was funded by a grant from the Max Bell Foundation to the first author (co-PIs AN and Tara Flanagan).

ACKNOWLEDGMENTS

We are grateful to Mélissa Ferland, Amanda Saxe, Janet Bang, and Jenny Wang for their assistance with testing, as well as Chloe Chow, Johanna Gruber, and Rachelle Doucet for work on transcription and detailed coding of these sometimes comical exchanges. We acknowledge the key contributions of Tara Flanagan and Keeley White to the McGill Transition Program for Adults with ASD. We thank the anonymous reviewers for their constructive input, and the participants who took part in this study.

REFERENCES

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th Edn*. Washington, DC: American Psychiatric Association.

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *J. Child Psychol. Psychiatry* 30, 285–298. doi: 10.1111/j.1469-7610.1989.tb00241.x

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Begeer, S., Malle, B. F., Nieuwland, M. S., and Keysar, B. (2010). Using theory of mind to represent and take part in social interactions: Comparing individuals with high-functioning autism and typically developing controls. *Eur. J. Dev. Psychol.* 7, 104–122. doi: 10.1080/17405620903024263

Bortfeld, H., and Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non—native speakers. *Discourse Process* 23, 119–147. doi: 10.1080/01638537709544986

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 41–57. doi: 10.1016/j.cognition.2011.05.011

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493. doi: 10.1037/0278-7393.22.6.1482

Brennan, S. E., and Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Top. Cogn. Sci.* 1, 274–291. doi: 10.1111/j.1756-8765.2009.01019.x

Brennan, S. E., Schuhmann, K. S., and Batres, K. M. (2013). "Entrainment on the move and in the lab: The walking around corpus," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (Austin, TX: Cognitive Science Society).

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *J. Mem. Lang.* 61, 171–190. doi: 10.1016/j.jml.2009.04.003

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Lang. Cogn. Process.* 27, 62–89. doi: 10.1080/01690965.2010.543363

Brown-Schmidt, S., and Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue Discourse* 2, 11–33. doi: 10.5087/dad.2011.102

Capps, L., Kehres, J., and Sigman, M. (1998). Conversational abilities among children with autism and children with developmental delays. *Autism* 2, 325–344. doi: 10.1177/1362361398024002

Clark, H. H. (1992). *Arenas of Language Use*. Chicago, IL: University of Chicago Press.

Clark, H. H., and Marshall, C. R. (1981). "Definite reference and mutual knowledge," in *Elements of Discourse Understanding*, eds A. K. Joshi, B. Webber, and I. A. Sag (Cambridge: Cambridge University Press), 10–63.

Clark, H. H., and Murphy, G. L. (1982). Audience design in meaning and reference. *Adv. Psychol.* 9, 287–299. doi: 10.1016/S0166-4115(09)60059-5

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Dahlgren, S., and Sandberg, A. D. (2008). Referential communication in children with autism spectrum disorder. *Autism* 12, 335–348. doi: 10.1177/1362361308091648

DE Marchena, A., and Eigsti, I.-M. (2016). The art of common ground: Emergence of a complex pragmatic language skill in adolescents with autism spectrum disorders. *J. Child Lang.* 43, 43–80. doi: 10.1017/S0305000915000070

de Villiers, J., Fine, J., Ginsberg, G., Vaccarella, L., and Szatmari, P. (2007). Brief report: A scale for rating conversational impairment in autism spectrum disorder. *J. Autism Dev. Disord.* 37, 1375–1380. doi: 10.1007/s10803-006-0264-1

Frith, U., and Happe, F. (1994). Autism: Beyond "theory of mind." *Cognition* 50, 115–132.

Garrod, S., and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* 27, 181–218.

Geurts, H. M., Bergh, S. F., and Ruzzano, L. (2014). Prepotent response inhibition and interference control in autism spectrum disorders: Two meta−Analyses. *Autism Res.* 7, 407–420. doi: 10.1002/aur.1369

Graham, S. A., Sedivy, J., and Khu, M. (2014). That's not what you said earlier: Preschoolers expect partners to be referentially consistent. *J. Child Lang.* 41, 34–50. doi: 10.1017/S0305000912000530

Hala, S., Pexman, P. M., and Glenwright, M. (2007). Priming the meaning of homographs in typically developing children and children with autism. *J. Autism Dev. Disord.* 37, 329–340. doi: 10.1007/s10803-006-0162-6

Horton, W. S., and Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition* 96, 127–142. doi: 10.1016/j.cognition.2004.07.001

Hus, V., and Lord, C. (2014). The autism diagnostic observation schedule, module 4: Revised algorithm and standardized severity scores. *J. Autism Dev. Disord.* 44, 1996–2012. doi: 10.1007/s10803-014-2080-3

Klin, A., Volkmar, F., and Sparrow, S. (1992). Autistic social dysfunction: Some limitations of the theory of mind hypothesis. *J. Child Psychol. Psychiatry* 33, 861–876. doi: 10.1111/j.1469-7610.1992.tb01961.x

Köymen, B., Schmerse, D., Lieven, E., and Tomasello, M. (2014). Young children create partner-specific referential pacts with peers. *Dev. Psychol.* 50, 2334–2342. doi: 10.1037/a0037837

Krauss, R. M. (1987). The role of the listener: Addressee influences on message formulation. *J. Lang. Soc. Psychol.* 6, 81–98. doi: 10.1177/0261927X8700600201

Krauss, R. M., and Glucksberg, S. (1969). The development of communication: Competence as a function of age. *Child Dev.* 40, 255–256. doi: 10.2307/1127172

Krauss, R. M., and Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychon. Sci.* 1, 113–114. doi: 10.3758/BF03342817

Krauss, R. M., and Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *J. Pers. Soc. Psychol.* 4, 343. doi: 10.1037/h0023705

Kronmüller, E., and Barr, D. J. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis. *J. Mem. Lang.* 56, 436–455. doi: 10.1016/j.jml.2006.05.002

Kronmüller, E., and Barr, D. J. (2015). Referential precedents in spoken language comprehension: A review and meta-analysis. *J. Mem. Lang.* 83, 1–19. doi: 10.1016/j.jml.2015.03.008

Kuhlen, A. K., and Brennan, S. E. (2013). Language in dialogue: When confederates might be hazardous to your health. *Psychon. Bull. Rev.* 20, 54–72. doi: 10.3758/s13423-012-0341-8

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., and Bishop, S. (2012). *Autism Diagnostic Observation Schedule, 2nd Edn (ADOS-2)*. Los Angeles, CA: Western Psychological Corporation.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101:676. doi: 10.1037/0033-295X.101.4.676

Matthews, D., Lieven, E., and Tomasello, M. (2010). What's in a manner of speaking? Children's sensitivity to partner-specific referential precedents. *Dev. psychol.* 46, 749. doi: 10.1037/a0019657

Metzing, C., and Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. *J. Mem. Lang.* 49, 201–213. doi: 10.1016/S0749-596X(03)00028-7

Mottron, L., Dawson, M., Soulières, I., Hubert, B., and Burack, J. (2006). Enhanced perceptual functioning in autism: An update, and eight principles of autistic perception. *J. Autism Dev. Disord.* 36, 27–43. doi: 10.1007/s10803-005-0040-7

Nadig, A., Lee, I., Singh, L., Bosshart, K., and Ozonoff, S. (2010). How does the topic of conversation affect verbal exchange and eye gaze? A comparison between typical development and high-functioning autism. *Neuropsychologia* 48, 2730–2739. doi: 10.1016/j.neuropsychologia.2010.05.020

Nadig, A., Vivanti, G., and Ozonoff, S. (2009). Adaptation of object descriptions to a partner under increasing communicative demands: A comparison of children with and without autism. *Autism Res.* 2, 334–347. doi: 10.1002/aur.102

Philofsky, A., Fidler, D. J., and Hepburn, S. (2007). Pragmatic language profiles of school-age children with autism spectrum disorders and Williams syndrome. *Am. J. Speech Lang. Pathol.* 16, 368–380. doi: 10.1044/1058-0360(2007/040)

Rutter, M., Bailey, A., and Lord, C. (2003). *SCQ. The Social Communication Questionnaire*. Torrance, CA: Western Psychological Services.

Schober, M. F., and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cogn. Psychol.* 21, 211–232. doi: 10.1016/0010-0285(89)90008-X

Shintel, H., and Keysar, B. (2007). You said it before and you'll say it again: Expectations of consistency in communication. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 357–369. doi: 10.1037/0278-7393.33.2.357

Shintel, H., and Keysar, B. (2009). Less is more: A minimalist account of joint action in communication. *Top. Cogn. Sci.* 1, 260–273. doi: 10.1111/j.1756-8765.2009.01018.x

Slocombe, K. E., Alvarez, I., Branigan, H. P., Jellema, T., Burnett, H. G., Fischer, A., et al. (2013). Linguistic alignment in adults with and without Asperger's syndrome. *J. Autism Dev. Disord.* 43, 1423–1436. doi: 10.1007/s10803-012-1698-2

Trueswell, J. C. and Tanenhaus, M. K. (1994). "Toward a lexicalist framework for constraint-based syntactic ambiguity resolution," in *Perspectives on Sentence Processing*, eds C. Clifton, L. Frazier, and K. Rayner (Mahwah, NJ: Lawrence Earlbaum Associates), 155–179.

Volden, J., and Lord, C. (1991). Neologisms and idiosyncratic language in autistic speakers. *J. Autism Dev. Disord.* 21, 109–130. doi: 10.1007/BF02284755

Volden, J., Mulcahy, R. F., and Holdgrafer, G. (1997). Pragmatic language disorder and perspective taking in autistic speakers. *Appl. Psycholingust.* 18, 181–198. doi: 10.1017/S0142716400009966

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale, 4th Edn (WAIS–IV)*. San Antonio, TX: NCS Pearson.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 20 November 2015
doi: 10.3389/fpsyg.2015.01787

EEG can Track the Time Course of Successful Reference Resolution in Small Visual Worlds

Christian Brodbeck[1,3]*, *Laura Gwilliams*[1,3] *and Liina Pylkkänen*[1,2,3]

[1] Department of Psychology, New York University, New York, NY, USA, [2] Department of Linguistics, New York University, New York, NY, USA, [3] NYUAD Institute, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Previous research has shown that language comprehenders resolve reference quickly and incrementally, but not much is known about the neural processes and representations that are involved. Studies of visual short-term memory suggest that access to the representation of an item from a previously seen display is associated with a negative evoked potential at posterior electrodes contralateral to the spatial location of that item in the display. In this paper we demonstrate that resolving the reference of a noun phrase in a recently seen visual display is associated with an event-related potential that is analogous to this effect. Our design was adapted from the visual world paradigm: in each trial, participants saw a display containing three simple objects, followed by a question about the objects, such as *Was the pink fish next to a boat?*, presented word by word. Questions differed in whether the color adjective allowed the reader to identify the referent of the noun phrase or not (i.e., whether one or more objects of the named color were present). Consistent with our hypothesis, we observed that reference resolution by the adjective was associated with a negative evoked potential at posterior electrodes contralateral to spatial location of the referent, starting approximately 333 ms after the onset of the adjective. The fact that the laterality of the effect depended upon the location of the referent within the display suggests that reference resolution in visual domains involves, at some level, a modality-specific representation. In addition, the effect gives us an estimate of the time course of processing from perception of the written word to the point at which its meaning is brought into correspondence with the referential domain.

Keywords: EEG/ERP, reference resolution, visual short-term memory, contralateral activity, language comprehension, reading

OPEN ACCESS

Edited by:
Emiel Krahmer,
Tilburg University, Netherlands

Reviewed by:
Olaf Hauk,
MRC Cognition and Brain Sciences Unit, UK
Vivien Mast,
University of Potsdam, Germany

****Correspondence:***
Christian Brodbeck
christianbrodbeck@nyu.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *01 August 2015*
Accepted: *06 November 2015*
Published: *20 November 2015*

Citation:
Brodbeck C, Gwilliams L and Pylkkänen L (2015) EEG can Track the Time Course of Successful Reference Resolution in Small Visual Worlds. Front. Psychol. 6:1787. doi: 10.3389/fpsyg.2015.01787

INTRODUCTION

Identifying the entities that individual expressions refer to is a fundamental prerequisite for understanding language in context. Even though EEG has been used widely to study language comprehension, so far no neural marker of successful reference resolution has been described. In this study we demonstrate that EEG can be used to track reference resolution by using visual displays as the referential domain. In this context, successful reference resolution is associated with an evoked potential known from research on visual short-term memory.

The cognitive basis of referential processing has been extensively studied with the so-called visual world paradigm (for a recent review, see Huettig et al., 2011). In these studies, participants typically look at a visual display while listening to instructions involving the display, and an eye tracker is used to determine what objects participants look at as the sentence unfolds. Results from visual world studies have underlined the centrality of referential processing for language comprehension by suggesting that the referential context can influence early syntactic parsing decisions (Tanenhaus et al., 1995; Spivey et al., 2002).

Concerning the process of reference resolution itself, visual world studies have suggested that it is fast and uses new information incrementally. When listeners followed spoken instructions such as "touch the starred yellow square," they fixated the referent shortly after hearing the word that allowed them to uniquely identify the referent, i.e., in an environment with only one starred item they fixated that item shortly after the word "starred," whereas in an environment with two starred items but only one of them yellow they fixated that item shortly after hearing "yellow" (Eberhard et al., 1995; Sedivy et al., 1999). Studies with more complex contexts have shown that eye movements in scenes are not just reactive to linguistic input but instead reflect listeners' predictions about upcoming referents, by, for example, fixating on a cake when hearing the verb "eat" (Altmann and Kamide, 1999, 2007; Kamide et al., 2003).

While eye tracking studies with the visual world paradigm have shed light on various aspects of reference resolution, not much is known about the time course of corresponding neural processes. Indirect evidence comes from a group of EEG studies which established that referential ambiguity is associated with a sustained negative evoked potential at frontal electrode sites, identified as "Nref" (reviewed by Nieuwland and Van Berkum, 2008). With serial visual presentation, referentially ambiguous determiner-noun phrases evoked an Nref around 300 ms after presentation of the noun (Van Berkum et al., 1999). A similar effect to referentially ambiguous pronouns had an onset around 400 ms (Nieuwland and Van Berkum, 2006). These results establish a time frame for referential processing by showing when the brain starts responding to referential ambiguity. There is some evidence that the Nref is specific to referential ambiguity, as pronoun resolution difficulty from sources other than referential ambiguity is not associated with an Nref (Van Berkum et al., 2007). Another relevant EEG study used a continuously presented visual world and auditory sentence stimuli, reporting a late central positive "P600" effect, commonly associated with syntactic violations and ensuing reanalysis, in a 500–800 ms time window when it became clear based on the visual world that a grammatically acceptable language fragment had to be reanalyzed as a less preferred construction (Knoeferle et al., 2008). While not directly reflecting referential processing, this still demonstrates an interaction between visual and linguistic information.

With the intention of establishing a neural marker of successful reference resolution for simple, unambiguous referential expressions, we sought to take advantage of the simplicity of the visual world paradigm. In its canonical form the visual world paradigm is not well suited for EEG data collection, where eye movements cause artifacts that overshadow brain signals. In order to overcome this problem, we modified the mode of presentation: In each trial, the visual world display was only shown for a short time and then replaced by a question presented centrally, word by word (see illustration in **Figure 1**). Participants' task was to focus on answering the question, using an internalized representation of the display hypothesized to reside in visual short term memory.

This paradigm allowed us to capitalize on a well-known effect from the literature on visual short-term memory: Directing attention to one side of the visual field is associated with a negative evoked potential at posterior electrodes contralateral to the side of attention (henceforth: "posterior contralateral negativity"). Originally described as an attention-dependent enhancement of the stimulus-dependent N2 component (e.g., Eason et al., 1969; Vanvoorhis and Hillyard, 1977; Heinze et al., 1990) this effect also occurs as a sustained posterior contralateral negativity when participants are instructed to maintain stimuli from only one side of a display in short term memory (e.g., Klaver et al., 1999; Vogel and Machizawa, 2004). Interestingly, a posterior contralateral negativity can also be observed in response to a centrally presented stimulus if the task requires relating it to an earlier, lateralized presentation (Gratton et al., 1997; Kuo et al., 2009; Dell'Acqua et al., 2010; Eimer and Kiss, 2010). These studies suggest that visual short-term memory traces are encoded and accessed in a topographic manner related to modality-specific neural pathways (cf. Klaver et al., 1999). Based on these results, we hypothesized that reference resolution in a referential domain held in visual short-term memory similarly entails access to modality-specific memory traces, which should manifest in a posterior contralateral negativity.

Some evidence concerning the involvement of the visual system in language processing comes from visual world studies in which the visual world display was shown only initially, and then removed and followed by a blank screen while participants listened to a sentence referring to objects in the display. Even on that blank screen, participants tended to fixate the previous location of the mentioned objects (Altmann, 2004; Altmann and Kamide, 2009). This was interpreted as reflecting access to an internal scene or event representation linking objects with their (prior) spatial location, and suggests that even in the absence of a concurrent visual display reference resolution in visual scenes held in short-term memory involves at some level access to a modality-specific visuo-spatial representation.

MATERIALS AND METHODS

Participants

We collected EEG data from 14 native speakers of English at New York University, Abu Dhabi. Data collection happened on a subset of participants taking part in a larger magnetoencephalography (MEG) study testing a different set of hypotheses. All participants had normal or corrected-to-normal vision, and none were colorblind or had known neurological abnormalities. Data from one subject were excluded because fewer than 50% of the trials remained after artifact rejection, leaving eight female and five male participants in the final analysis (mean age 24.3, range 18–38 years). All participant had acquired English as their first language, but three of the 13 grew up speaking at least one other language. The protocol was approved by the Institutional Review Board of NYU Abu Dhabi, and all participants provided written consent before beginning the experiment.

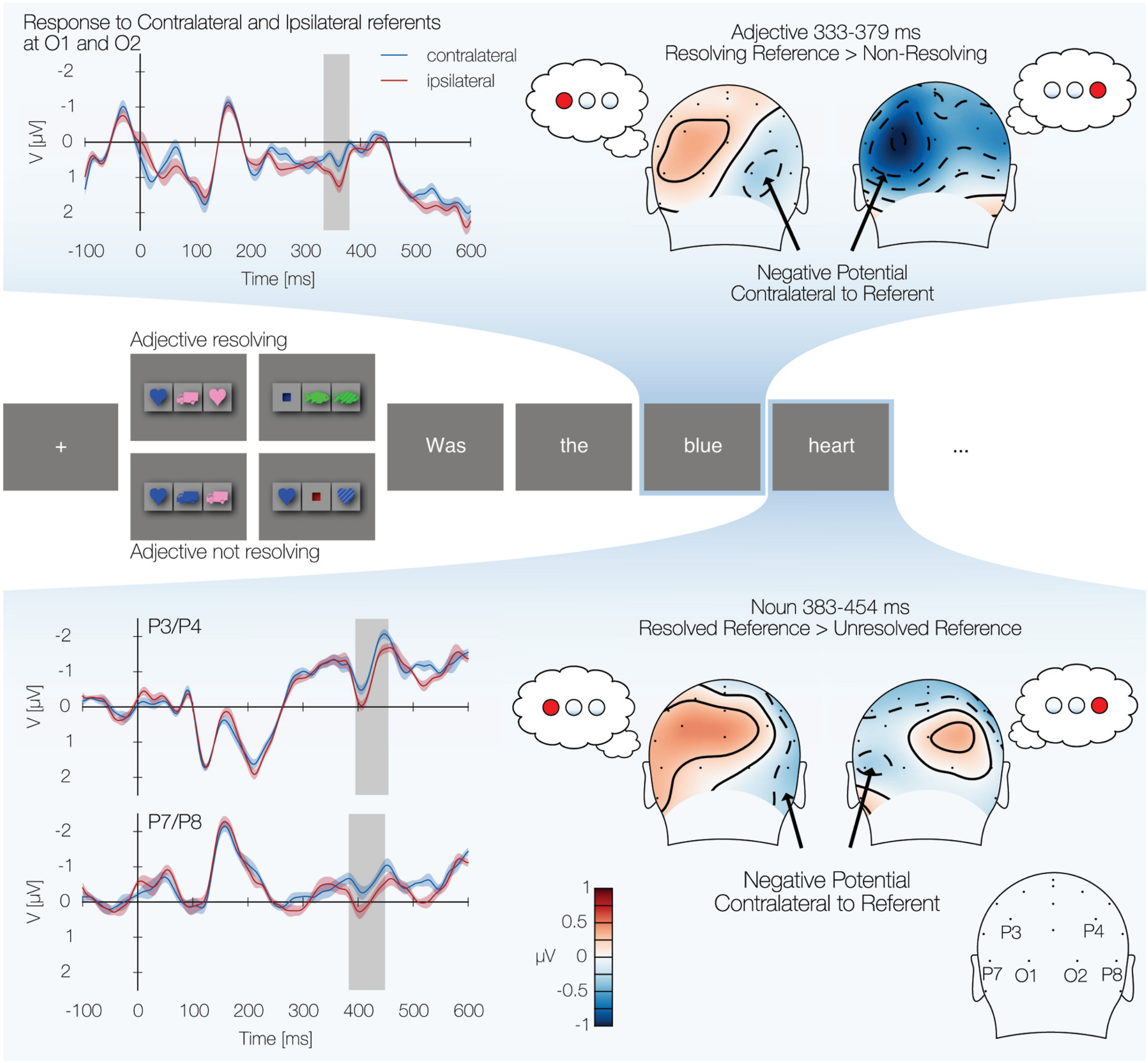


FIGURE 1 | Posterior negativity contralateral to the referents of adjectives and nouns. Top panel: the left side shows the response to the adjective at electrodes O1 and O2, with responses grouped according to whether reference was resolved to the side contralateral or ipsilateral to the sensor. The shaded area indicates the region in which the response to contralateral referents was significantly more negative than the response to ipsilateral referents. The topographies on the right show difference maps of the average voltage during the time window established by the cluster, seen from the back of the head; for referents on the left and right, the average potential is plotted for resolving adjectives to that side minus all non-resolving adjectives. **Middle panel:** illustration of the paradigm, each rectangle representing a computer screen. Each trial started with a fixation cross. Next a visual world display was presented (the four different displays shown illustrate different experimental conditions). After the display disappeared, a question about the display was presented word by word. **Bottom panel:** response to nouns at which either the adjective had already resolved reference, or where the noun itself resolved reference. Plots are analogous to the top panel. Topographies are contrasted to the response to the nouns at which reference had not been resolved yet.

Design and Stimuli

Each trial consisted of a visual world display and a corresponding question (see **Figure 1**, middle panel for an illustration). The visual world display and content words were presented for 300 ms with an interstimulus interval (ISI) of 300 ms, whereas short function words like “was,” “the,” “on,” and “in” were presented for 200 ms with an ISI of 200 ms. The last word of the question stayed on the screen until participants gave a yes/no answer by pressing one of two buttons.

Each visual world display contained three horizontally aligned objects. Objects were constructed on the basis of six colors (blue, brown, green, pink, red, white) and five shapes (boat, fish, heart, star, truck). There were two kinds of displays (a manipulation that was mainly of interest for the MEG study and is not discussed further here): The first, simpler kind contained one pair of objects that shared color and another pair that shared shapes, with all shapes visible. The second, more complex kind of display contained one object of unique color, but with its shape occluded, and two additional objects, both of the same color and shape, but with differing patterns.

All questions began with "Was the [color-adjective] [shape-noun] . . .", and asked about the absolute or relative position of one of the items in the display, for example "Was the pink fish next to a boat?" or "Was the blue heart in the middle?". Different kinds of questions were used to discourage participants from relying on specific strategies focusing on particular aspects of the visual displays. The correct answer was "yes" in half of the trials and "no" in the other half. The color adjectives and shape identifying nouns used in referential expressions (*blue*, *brown*, *green*, *pink*, *red*, *white*; *boat*, *fish*, *heart*, *star*, *truck*) were all common words with SUBTL frequencies between 28.5 and 244.2 per million words (Brysbaert and New, 2009) and mean lexical decision latencies between 523 and 653 ms according to the English Lexicon Project (Balota et al., 2007).

The complete design included the factors reference (adjective resolving vs. adjective non-resolving), display kind (simple or complex), location of the referent (left, middle, right, with middle treated as fillers for the purpose of this analysis), color (six levels) and shape (five levels) for a total of $2 \times 2 \times 3 \times 6 \times 5 = 360$ trials. Each participant saw the same 360 trials, but the order was randomized for each experimental session. Thus, for each possible location of the referent (left, middle, right), there were 60 trials in which reference was resolved by the adjective (top left and top right displays in **Figure 1**). In 30 trials the noun resolved reference (bottom left display in **Figure 1**), and in another 30 trials reference was resolved by a prepositional phrase following the noun (bottom right display in **Figure 1**).

Procedure

Participants were given instructions on the reference task and allowed to practice using sample trials until they felt comfortable performing the task. Recordings took place concurrent with MEG recordings, inside a magnetically shielded MEG acquisition chamber. Participants lay in a supine position and stimuli were projected onto a horizontal screen at comfortable viewing distance. Participants were instructed to blink as little as possible during the presentation of the stimuli. They were told that if they needed a break they could withhold their yes/no response at the end of a trial until they felt comfortable to continue. In regular intervals throughout the experiment they were informed of the progress in the experiment with a text display and had the opportunity to take a short, self-terminated break. Stimuli were presented with MATLAB using psychtoolbox[1] and ptbwrapper[2]. On average, an experimental session took 42 min from first to last trial (without setup).

Data were recorded from 31 EEG and 3 EOG electrodes attached to an elastic cap at standard positions in the international 10–20 system (EasyCap GmbH, Germany) at a sampling rate of 1000 Hz. Impedances were kept below 10 kΩ. The ground was located at the AFZ electrode position, and recordings were referenced to the left mastoid electrode. The signal was amplified with a BrainVision Brain Amp Standard amplifier.

Data Analysis

Data were pre-processed and analyzed with MNE-Python (Gramfort et al., 2013, 2014) and Eelbrain[3]. Raw data were band-pass filtered offline between 0.1 and 40 Hz. We extracted epochs from −100 to 600 ms relative to the onset of the adjectives. Epochs containing artifacts were excluded from further analysis, and individual channels containing noise were interpolated. Artifact rejection proceeded with automatic rejection of epochs with a signal exceeding an absolute 7.5 μV threshold, followed by adjustment based on visual inspection. If individual channels exhibited signal at abnormal amplitude independently of neighboring channels, the signal at aberrant channels was interpolated using spherical spline interpolation from the remaining channels instead of rejecting the whole epoch. Epochs were baseline corrected using the 100 ms pre-adjective period and re-referenced to the average of the two mastoid electrodes.

The statistical analysis focused on the lateral posterior electrodes, O1, O2, P3, P4, P7, and P8. These electrodes represent the lateral posterior part of the head in our electrode layout, where N2pc and contralateral delay activity are most reliably observed (see literature cited in the introduction). For each subject and electrode pair (O1/O2, P3/P4, and P7/P8) we computed one wave form for adjectives resolving reference to the side contralateral to the electrode, and a second waveform for adjectives resolving reference to the side ipsilateral with the electrode. We then tested the hypothesis that the contralateral signal was more negative than the ipsilateral signal with temporal cluster-permutation tests based on one-tailed t-tests (see Nichols and Holmes, 2002; Maris and Oostenveld, 2007). We performed the test on a time window from 200 to 500 ms after adjective onset. The beginning of this time window was based on the onset of the posterior contralateral negativity in visual short-term memory studies around 200–250 ms (e.g., Dell'Acqua et al., 2010) and the offset was based on when people moved their eyes to the referent identified by an adjective in visual world studies (Eberhard et al., 1995). For each time point (at a resolution of 1000 Hz) we calculated a related-samples t-value. We then formed clusters based on contiguous regions of t-values greater or equal to a value equivalent with an uncorrected one-tailed p-value of 0.05. For each cluster we calculated the cluster mass (i.e., the sum of the t-values making up the cluster). We then repeated this procedure with all 8191 possible permutations of the data (with 13 participants there are 2^{13}–1 possible ways of switching condition labels within subjects) and extracted for each permutation the maximum cluster mass value. The distribution of these values provides a non-parametric estimate of the expected distribution of the maximum cluster mass statistic under the null hypothesis. This distribution was used to assign to each cluster in the actual data a p-value corrected for multiple comparisons across time, by locating the cluster's mass on the distribution. Since we performed this procedure at three electrode pairs we multiplied all resulting p-values by 3.

We analyzed the response to nouns with the same procedure with epochs extracted around the onset of presentation of the

[1] psychtoolbox.org

[2] code.google.com/p/ptbwrapper

[3] https://pythonhosted.org/eelbrain

nouns. Due to the low number of trials in which the noun actually resolved reference, we performed a cluster-permutation analysis over trials in which reference was resolved by the adjective or the noun, and used the time window identified by this analysis for targeted *post hoc* tests in the sub-conditions.

RESULTS

Behavioral

On average, participants answered 87.6% of the questions correctly, ranging from 77.2 to 97.8% correct.

EEG Response to Adjectives

After artifact rejection, an average of 55.2 trials per participant per condition (reference resolution to the left, reference resolution to the right) remained of a total of 60 possible. A significant cluster in which the signal contralateral to the referent was more negative than the ipsilateral signal was found in the O1/O2 sensor pair (333–379 ms, $p = 0.019$ Bonferroni-corrected). **Figure 1** (top panel, left side) shows the contralateral and ipsilateral response to the adjectives, with the time window of the significant cluster shaded. The accompanying topographic plots show the average potential, during the time window identified by the cluster, for reference resolution toward the object on the left (or the right) side of the display minus the average of the non-resolving adjectives, illustrating the presence of the posterior contralateral negativity.

EEG Response to Nouns

Our initial analysis of the response to the noun included all trials in which after reading the noun, participants could know the location of the referent. This included trials in which reference was resolved by the adjective and trials in which reference was resolved by the noun. This yielded an average of 82.0 trials per referent location condition (referent on the left vs. referent on the right) out of 90 possible. In this combined response, we found a significant posterior contralateral negativity at the P3/P4 electrode pair (395–454 ms, $p = 0.025$ Bonferroni-corrected) and at the P7/P8 electrode pair (383–449 ms, $p = 0.027$ Bonferroni-corrected). **Figure 1** (lower panel) shows the contralateral and ipsilateral responses at P3/P4 and P7/P8. To illustrate the topography of the effects, the figure shows topographic maps in which the response to nouns with known referents on the left or right side of the display is compared to the response to nouns at which the location of the referent was still unknown.

For follow-up analysis in the sub-conditions we calculated the average of the P3/P4 and P7/P8 sensor pairs in the time window 395–449 ms, in which the two clusters overlapped. This analysis confirmed the recurrence of a posterior contralateral negativity in the response to nouns in the adjective resolving condition [difference = -5.34 μV, $t(12) = 3.50$, $p_{\text{one-tailed}} = 0.002$]. In the noun-resolving condition, in which the nouns followed adjectives that were compatible with two objects, the difference was in the expected direction, but did not reach significance [difference = -2.98 μV, $t(12) = 1.12$, $p_{\text{one-tailed}} = 0.14$]. This result begs the question whether we simply lacked the power to detect the response to resolving nouns due to the low number of trials in this condition, or whether the response to resolving nouns was indeed different form the response to non-resolving nouns. This latter hypothesis would predict a significant difference between the contralateral negativity in the response to resolving and non-resolving nouns; however, a related measures t-test indicated that this was not the case [difference = 2.36 μV, $t(12) = 0.72$, $p = 0.49$]. Therefore, our data do not let us draw a conclusion about the response to reference-resolving nouns.

One possible explanation for a contralateral response to nouns after reference-resolving adjectives is that on some trials, readers failed to resolve reference on the adjective, even though this would have been possible, and caught up by resolving reference when they read the noun. This line of reasoning would suggest a negative relationship, trial by trial, between the contralateral negativity on the adjective and the contralateral negativity on the noun. In order to test whether this was the case we calculated, for each subject, within the adjective-resolving trials, the correlation coefficient between the contralateral response to the adjectives and the contralateral response to the nouns. The contralateral effects were quantified as the contralateral minus ipsilateral difference of the average of the time points and sensors involved in the significant clusters described above. A one sample t-test indicated that these correlation coefficients were not reliably different from 0 across subjects [mean $r = -0.01$, $t(12) = -0.53$, $p = 0.61$]. This indicates that the contralateral response to the nouns in cases where the adjective had already resolved reference was not contingent upon the absence of a contralateral response to the adjective, i.e., that participants tended to show a contralateral response to both words.

DISCUSSION

We investigated whether a posterior contralateral EEG response previously observed in visual short-term memory tasks is also present when linguistic expressions refer to objects held in visual short-term memory. We analyzed the response to visually presented adjective–noun phrases, embedded in a natural context of questions about visual displays. As predicted, we found that reference resolution was associated with a negative evoked potential at posterior electrodes contralateral to the site of the referent.

When adjectives resolved reference (i.e., color adjectives in contexts where only one object had that color) they were associated with a posterior negativity contralateral to the referent, starting 333 ms after presentation of the adjective. Importantly, the conditions we compared involved the same adjectives; what differed between conditions was the location of the referent picked out by the adjectives. The fact that the signal reflected the spatial position of the referent within the referential domain strongly implies that it was due to a process associated with reference resolution, for which that location mattered, rather than a process that is independent of the location of the referent (such as, for example, a cloze probability effect).

Our results leave open the exact nature of the process that produces the observed effect. On the one hand, the effect does not necessarily reflect commitment to one specific object as the referent for the given linguistic expression; it would also be compatible with an evaluation process supporting reference resolution, for example by activating those spatial locations in a visual short-term memory representation that include the color named by the adjective. On the other hand, an alternative possibility is that reference is resolved in an abstract, modality-general representation, and the visual representation is only accessed once the referent is found in the abstract representation. These questions are open to future research. However, the presence of a an EEG signal that reflects the spatial position of the referent does suggest that contact between the semantics of the word (color adjective) and the referential domain has been established, and that the adjective leads readers to activate the portion of the referential domain that contained the item with the corresponding color.

The response to adjectives reflecting a reference resolution process is consistent with findings from visual world studies which showed that reference resolution is incremental, i.e., that language comprehenders use each incoming word of a referential expression to constrain the set of possible referents (Eberhard et al., 1995). In addition, while visual world studies used spoken language input, our results extend this observation to the context of reading.

For nouns, the posterior contralateral negativity started around 383 ms and was significant even when the noun merely occurred as the head of an expression for which reference had already been resolved. This indicates that even when readers had supposedly already identified the referent, they still reactivated the corresponding representation when processing the noun. This finding could be relevant for models of the comprehension of overspecified referential expressions (for an overview, see Gatt et al., 2014). Language producers frequently overspecify referential expressions, in particular involving colors, for example, using *the blue heart* in a context in which *the heart* would have been sufficient to distinguish the referent from its competitors (Pechmann, 1989). In simple contexts, overspecified expressions have been argued to speed up (e.g., Arts et al., 2011) or slow down comprehension (e.g., Engelhardt et al., 2011). Our results suggest that our participants processed redundant information by reactivating the referent they had already identified through the adjective. This might indicate that, regardless of processing speed, overspecification is associated with more robust comprehension. For example, participants might have reactivated the representation of the referent when reading the noun to check their initial interpretation after the adjective. This increase in robustness could be particularly relevant in real life referential domains, which are often more complex and less constrained than experimental stimuli. Indeed, it has been shown that overspecification can significantly simplify the referential search in certain more complex referential domains (Paraboni et al., 2007; Paraboni and van Deemter, 2013).

While the latency difference between adjectives and nouns is suggestive, it would seem premature to draw definite conclusions, especially since our design did not include enough trials on which the noun resolved reference.

Our results put the time point at which readers identify a referent's location in response to a visually presented content word around 350 ms. Even though the effect we described is of a quite different nature, it converges with studies of referential ambiguity (e.g., Van Berkum et al., 1999) to place the time point at which linguistic input starts interacting with the referential domain between 300 and 400 ms.

More broadly, the observation of a posterior negativity contralateral to the referent of a linguistic expression suggests that people use the same or similar memory systems when understanding language as in non-linguistic visual short-term memory tasks. If this interpretation is correct, it suggests that people use domain-specific, non-linguistic representations when comprehending referential expressions. This observation fits well into the broader context of research suggesting that language comprehension engages domain-specific cognitive mechanisms to process linguistic meaning (e.g., Zwaan et al., 2002). Our results suggest that it is possible to track the mind's eye looking at a visual memory when reading about it.

ACKNOWLEDGMENTS

We thank the reviewers, Olaf Hauk and Vivien Mast, and the editor, Emiel Krahmer, for their helpful comments and suggestions. This research was supported by the National Science Foundation Grant BCS-1221723 (LP) and grant G1001 from the NYUAD Institute, New York University Abu Dhabi (LP).

REFERENCES

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the 'blank screen paradigm'. *Cognition* 93, B79–B87. doi: 10.1016/j.cognition.2004.02.005

Altmann, G. T. M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* 73, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., and Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: linking anticipatory (and other) eye movements to linguistic processing. *J. Mem. Lang.* 57, 502–518.

Altmann, G. T. M., and Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: eye movements and mental representation. *Cognition* 111, 55–71. doi: 10.1016/j.cognition.2008.12.005

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011). Overspecification facilitates object identification. *J. Pragmat.* 43, 361–374. doi: 10.1016/j.pragma.2010.07.013

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behav. Res. Methods* 39, 445–459. doi: 10.3758/BF03193014

Brysbaert, M., and New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* 41, 977–990. doi: 10.3758/BRM.41.4.977

Dell'Acqua, R., Sessa, P., Toffanin, P., Luria, R., and Jolicoeur, P. (2010). Orienting attention to objects in visual short-term memory. *Neuropsychologia* 48, 419–428. doi: 10.1016/j.neuropsychologia.2009.09.033

Eason, R. G., Harter, M. R., and White, C. T. (1969). Effects of attention and arousal on visually evoked cortical potentials and reaction time in man. *Physiol. Behav.* 4, 283–289. doi: 10.1016/0031-9384(69)90176-0

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *J. Psycholinguist. Res.* 24, 409–436. doi: 10.1007/BF02143160

Eimer, M., and Kiss, M. (2010). An electrophysiological measure of access to representations in visual working memory. *Psychophysiology* 47, 197–200. doi: 10.1111/j.1469-8986.2009.00879.x

Engelhardt, P. E., Bariş Demiral, Ş., and Ferreira, F. (2011). Over-specified referring expressions impair comprehension: an ERP study. *Brain Cogn.* 77, 304–314. doi: 10.1016/j.bandc.2011.07.004

Gatt, A., Krahmer, E., van Deemter, K., and van Gompel, R. P. G. (2014). Models and empirical data for the production of referring expressions. *Lang. Cogn. Neurosci.* 29, 899–911. doi: 10.1080/23273798.2014.933242

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front Neurosci.* 7:267. doi: 10.3389/fnins.2013.00267

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE software for processing MEG and EEG data. *Neuroimage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Gratton, G., Corballis, P. M., and Jain, S. (1997). Hemispheric organization of visual memories. *J. Cogn. Neurosci.* 9, 92–104. doi: 10.1162/jocn.1997.9.1.92

Heinze, H. J., Luck, S. J., Mangun, G. R., and Hillyard, S. A. (1990). Visual event-related potentials index focused attention within bilateral stimulus arrays. 1. Evidence for early selection. *Electroencephalogr. Clin. Neurophysiol.* 75, 511–527. doi: 10.1016/0013-4694(90)90138-A

Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: a review and critical evaluation. *Acta Psychol.* 137, 151–171. doi: 10.1016/j.actpsy.2010.11.003

Kamide, Y., Altmann, G. T. M., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *J. Mem. Lang.* 49, 133–156. doi: 10.1016/S0749-596X(03)00023-8

Klaver, P., Talsma, D., Wijers, A. A., Heinze, H. J., and Mulder, G. (1999). An event-related brain potential correlate of visual short-term memory. *Neuroreport* 10, 2001–2005. doi: 10.1097/00001756-199907130-00002

Knoeferle, P., Habets, B., Crocker, M. W., and Munte, T. F. (2008). Visual scenes trigger immediate syntactic reanalysis: evidence from ERPs during situated spoken comprehension. *Cereb. Cortex* 18, 789–795. doi: 10.1093/cercor/bhm121

Kuo, B.-C., Rao, A., Lepsien, J., and Nobre, A. C. (2009). Searching for targets within the spatial layout of visual short-term memory. *J. Neurosci.* 29, 8032–8038. doi: 10.1523/JNEUROSCI.0952-09.2009

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25. doi: 10.1002/hbm.1058

Nieuwland, M. S., and Van Berkum, J. J. A. (2006). Individual differences and contextual bias in pronoun resolution: evidence from ERPs. *Brain Res.* 1118, 155–167. doi: 10.1016/j.brainres.2006.08.022

Nieuwland, M. S., and Van Berkum, J. J. A. (2008). The neurocognition of referential ambiguity in language comprehension. *Lang. Linguist. Compass* 2, 603–630. doi: 10.1111/j.1749-818X.2008.00070.x

Paraboni, I., and van Deemter, K. (2013). Reference and the facilitation of search in spatial domains. *Lang. Cogn. Neurosci.* 29, 1002–1017. doi: 10.1080/01690965.2013.805796

Paraboni, I., van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: making referents easy to identify. *Comput. Linguist.* 33, 229–254. doi: 10.1162/coli.2007.33.2.229

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 89. doi: 10.1515/ling.1989.27.1.89

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition* 71, 109–147. doi: 10.1016/S0010-0277(99)00025-6

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., and Sedivy, J. C. (2002). Eye movements and spoken language comprehension: effects of visual context on syntactic ambiguity resolution. *Cogn. Psychol.* 45, 447–481. doi: 10.1016/S0010-0285(02)00503-0

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Van Berkum, J. J. A., Brown, C. M., and Hagoort, P. (1999). Early referential context effects in sentence processing: evidence from event-related brain potentials. *J. Mem. Lang.* 41, 147–182. doi: 10.1006/jmla.1999.2641

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., and Nieuwland, M. S. (2007). Establishing reference in language comprehension: an electrophysiological perspective. *Brain Res.* 1146, 158–171. doi: 10.1016/j.brainres.2006.06.091

Vanvoorhis, S., and Hillyard, S. A. (1977). Visual evoked-potentials and selective attention to points in space. *Percept. Psychophys.* 22, 54–62. doi: 10.3758/BF03206080

Vogel, E. K., and Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature* 428, 748–751. doi: 10.1038/nature02447

Zwaan, R. A., Stanfield, R. A., and Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychol. Sci.* 13, 168–171. doi: 10.1111/1467-9280.00430

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 January 2016
doi: 10.3389/fpsyg.2015.02035

Talker-Specific Generalization of Pragmatic Inferences based on Under- and Over-Informative Prenominal Adjective Use

Amanda Pogue[1]*, *Chigusa Kurumada*[1] *and Michael K. Tanenhaus*[1,2]

[1] *Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, USA,* [2] *Department of Linguistics, University of Rochester, Rochester, NY, USA*

According to Grice's (1975) Maxim of Quantity, rational talkers formulate their utterances to be as economical as possible while conveying all necessary information. Naturally produced referential expressions, however, often contain more or less information than what is predicted to be optimal given a rational speaker model. How do listeners cope with these variations in the linguistic input? We argue that listeners navigate the variability in referential resolution by calibrating their expectations for the amount of linguistic signal to be expended for a certain meaning and by doing so in a context- or a talker-specific manner. Focusing on talker-specificity, we present four experiments. We first establish that speakers will generalize information from a single pair of adjectives to unseen adjectives in a speaker-specific manner (Experiment 1). Initially focusing on exposure to underspecified utterances, Experiment 2 examines: (a) the dimension of generalization; (b) effects of the strength of the evidence (implicit or explicit); and (c) individual differences in dimensions of generalization. Experiments 3 and 4 ask parallel questions for exposure to over-specified utterances, where we predict more conservative generalization because, in spontaneous utterances, talkers are more likely to over-modify than under-modify.

Keywords: sentence processing, adaptation, generalization, pragmatics, informativity, referential expressions

OPEN ACCESS

Edited by:
Emiel Krahmer,
Tilburg University, Netherlands

Reviewed by:
Martijn Goudbeek,
Tilburg University, Netherlands
Catherine Davies,
University of Leeds, UK

Correspondence:
Amanda Pogue
apogue@mail.bcs.rochester.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *28 August 2015*
Accepted: *21 December 2015*
Published: *20 January 2016*

Citation:
Pogue A, Kurumada C and Tanenhaus MK (2016) Talker-Specific Generalization of Pragmatic Inferences based on Under- and Over-Informative Prenominal Adjective Use. Front. Psychol. 6:2035. doi: 10.3389/fpsyg.2015.02035

INTRODUCTION

A key feature of human language is that there are many-to-many mappings between referents and linguistic expressions. A pet dog can be referred to by many expressions (e.g., *the dog*, *Charlie*, *he*, or *my friend*) whereas the expression *the dog* can be used to refer to a real dog, a toy dog, or a contemptible person. Referential expressions can also be made arbitrarily long (e.g., *the big dog*, *the big brown dog*, *the big brown furry dog, etc.*). One long-standing issue in psycholinguistic research is how language users map a referential expression onto an intended referent with the speed and accuracy evidenced in real time language use (e.g., Altmann and Steedman, 1988; Eberhard et al., l995; Allopenna et al., 1998; Arnold et al., 2000; Brown-Schmidt and Hanna, 2011).

One influential hypothesis is that listeners cope with this mapping problem by assuming that speakers behave rationally, formulating their utterances to be as economical as possible while conveying all necessary information (Grice, 1975). Hereafter, we call this the rational-speaker model. For example, a rational speaker is more likely to use a pre-nominal scalar adjective (e.g., *the big dog*) when there is a complement (contrast) set of referents of the same semantic type (e.g., a big and one or more small dogs) in the same context (Sedivy et al., 1999; Davies and Katsos, 2009). By assuming a rational model of the speaker, listeners can make predictions about the referring expression that maximize the informativity of a linguistic element, where informativity is defined as the amount of uncertainty that is reduced by the element given the set of plausible referents in the current referential domain (Frank and Goodman, 2012). Frank and Goodman (2012) tested the informativity hypothesis using a simple language game. With three geometrical shapes with two shape features and two colors (e.g., a blue square, a blue circle, and a green square), comprehenders were asked to pick the referent that best matched a single word description (e.g., *blue* or *square*). A rational language user model predicts that when given *blue* participants should most frequently choose the blue square rather than the blue circle. This is because if the talker had meant the blue circle, she should have used the more informative (unambiguous) description *circle*. The results confirmed this prediction.

Real-time processing of prenominal adjectives is also influenced by the assumption that the speaker is formulating her utterances to efficiently pick out a referent given contextually salient contrast sets. In a visual world study (Tanenhaus et al., 1995), Sedivy et al. (1999) used spoken instructions such as “Pick up the tall glass” in a visual workspace with a tall glass and a short glass (which form a contrast set), a tall pitcher and an unrelated object (e.g., a key). A rational speaker would use the adjective “tall” to refer to the glass, which is a member of a contrast set, and not the tall pitcher. If listeners use the context and make this inference in real-time, as they hear the adjective, “tall,” they should begin to look at the tall glass. This is just the result reported by Sedivy et al. (also see Hanna et al., 2003; Heller et al., 2008; Wolter et al., 2011).

Although these results are consistent with a rational model of reference generation and understanding, some researchers have questioned whether a rational model will scale up to account for interlocutors’ behavior in everyday language use. Spontaneously produced referential expressions often include information that would be superfluous under the assumption that the speaker should only provide necessary and sufficient information. For example, spontaneous utterances often contain prenominal modifiers that are not necessary for identifying a unique referent (Deutsch and Pechmann, 1982; Sonnenschein, 1984; Pechmann, 1989; Belke, 2006; Engelhardt et al., 2006; see also Koolen et al., 2011). For instance, 30% of speakers used superfluous adjectives in a production study in Engelhardt et al. (2006) and 50% in Nadig and Sedivy (2002).

Conversely, interlocutors frequently under-specify in highly specific circumstances. In Brown-Schmidt and Tanenhaus (2008), interlocutors were tasked with rearranging blocks on puzzle boards. Areas in the workspace were divided into sub-regions. More than 50% of the referential expressions were underspecified with respect to potential referents in the relevant sub-region. Nonetheless, these utterances were seamlessly interpreted by the listener. Analyses showed that underspecified utterances only occurred when the alternatives were unlikely to be the intended referent given the local task constraints. For example a speaker might say, “put it above the red block,” when there were two red blocks but only one had a free space above it.

In sum, in relatively simple situations, like those typically examined in psycholinguistic studies, talkers often over-specify. In contrast, in more complex situations with richly structured discourse context, talkers frequently under-specify. For purposes of the present work, we will be focusing on situations where over-specification in the form of “redundant” prenominal adjectives is quite common and under-specification is relatively infrequent.

How can we reconcile the ubiquitous over-specification in these situations with the evidence that listeners seem to assume that a prenominal adjective is included to form a maximally informative utterance with respect to the context? One possibility is that the rational assumption is only one of many relevant factors that the talker and the listener take into account, rather than a strong determinant of reference generation and understanding. For example, in an interactive communication game, Engelhardt et al. (2006) reported comprehenders’ asymmetrical reactions to over- and under-modifying expressions. Comprehenders judged an ambiguous, under-specifying, expression in the presence of more than one plausible candidate to be less than optimal. However, they did not seem to draw additional inferences from superfluous, over-specifying, descriptors (see also Davies and Katsos, 2010, for evidence of asymmetrical penalization of over- vs. under-modified expressions by adults and children in a non-interactive task). Based on these asymmetrical findings between under- and over-informative utterances, Engelhardt et al. (2006, p. 572) concluded, “people are only moderately Gricean.”

Before adopting this conclusion, there is another approach that we believe is worth exploring. This approach is motivated, in part, by work that reevaluates what it means to be *rational* in decision-making. In a seminal line of research, Tversky and Kahneman (1974) documented ways in which human agents systematically deviate from the rational models widely assumed within economics. They proposed that people rely on heuristics, such as availability, similarity and representativeness that can result in fallacies leading to non-rational, or non-logical, decision-making under many circumstances. One such case is the “conjunction fallacy” where given a scenario about, Linda, a college-educated woman who cares deeply about social issues, participants will rate the likelihood the Linda is *bank teller and a feminist* as greater than the likelihood that she is *a bank teller* (Tversky and Kahneman, 1983). This clearly violates a basic rule of logic and probability—a conjunction cannot contain more members than either of its conjuncts. These fallacies were therefore taken to suggest that human agents are not rational in their decision-making behaviors.

However, the same evidence can be viewed as consistent with the hypothesis that participants are behaving according to basic assumptions about the rationality of language users. One of the assumptions is relevance in information. In Grice’s terms, “Our talk exchanges do not normally consist of a succession of disconnected remarks, *and would not be rational if they did*” (Grice, 1975, p. 45, emphasis original). Based on this assumption, when the talker provides certain information (e.g., *Linda is a feminist*), the listener infers that she must have had a reason to do so with respect to the goal of achieving successful communication. Rationality, in this sense, manifests itself in the general tendency for language users to engage in goal-directed acts of communication even in the simple task used by Tversky and Kahneman rather than simply treating the scenario as an abstract logical problem (Hertwig and Gigerenzer, 1999; also see Oaksford and Chater, 2001 for a similar approach applied to other decision problems). Thus what might appear to be departures from rationality are in fact grounded in principled behaviors that overall lead to more successful communication[1].

When we apply this perspective to reference generation and comprehension, it seems plausible that what we might view as departures from the rational-speaker model could, in fact, be fully consistent with a rational perspective. Let us assume that one of the most prominent goals of linguistic communication is to successfully convey intended messages and that this communication takes place through a noisy channel. It is essential, then, for the speaker to provide listeners with sufficient information while taking into account the likely possibility that some information will be lost due to noise in the production and comprehension systems (e.g., Aylett and Turk, 2004; Levy and Jaeger, 2007; Jaeger, 2010; Gibson et al., 2013). In particular, early in an interaction, interlocutors are likely to have uncertainty about the relevant context that bears on the current interaction and the degree to which they have shared goals and experience, etc. There is also variability in how well-different talkers and listeners take into account each other's perspective, individual differences along dimensions, such as spatial ability, and differences in speech style (e.g., the degree to which abrupt utterances are considered impolite). Given these considerations, it can be *rational* to provide more information than what is minimally required, rather than trying to estimate what degree of specification is optimal. This tendency is likely strengthened in non-interactive tasks in which a talker cannot negotiate with her listener during the interaction.

Indeed, there is evidence that listeners can take into account such communicative considerations from the speaker's perspective. Davies and Katsos (2009) proposed that the higher tolerance for over-informative expressions in Engelhardt et al. (2006) arises because these expressions can plausibly be attributable to communicative reasons (e.g., an extra effort for avoiding ambiguity). When the redundancy is unlikely to benefit communication, comprehenders found the over-informative utterances to be sub-optimal just as they do for under-informative utterances. Davies and Katsos' (2009) results suggest that listeners do not simply judge whether an expression is over-informative given a referent, but they reason about why the speaker produced an additional element with respect to the goal of successful communication. As conversation unfolds and as interlocutors have an increasingly coordinated construal of common ground, expectations for referring expressions are also tightened in a talker- and context-specific manner (Metzing and Brennan, 2003; Brown-Schmidt et al., 2015; Kronmüller and Barr, 2015). As a result, what might appear to be an ambiguous referring expression becomes fully informative for interlocutors, allowing them to communicate more efficiently (Brown-Schmidt and Tanenhaus, 2008).

From this perspective, in contrast to Engelhardt et al.'s (2006) proposal, we hypothesize that variations found in referring expressions reflect rational principles for maximizing overall communicative success under uncertain conditions. We posit that listeners assume that talkers are generally Gricean, rather than only sometimes Gricean. Crucially, our framework assumes that (1) listeners expect talkers to vary in their choices of referential expressions and that (2) listeners constantly adapt their expectations about how much linguistic information a particular talker might provide to convey a particular referential intention. This allows listeners to navigate the variability in referring expressions to arrive at the intended referent.

As a first step in developing this approach, the current paper tests whether and, if so, how listeners adapt their referential expectations in simple communicative contexts similar to those used in many other psycholinguistic studies discussed above. In particular, we ask whether listeners adapt their expectations in a talker-specific fashion. This question is motivated by Grodner and Sedivy's (2011) demonstration that listeners discount linguistic evidence for contrastive inference when they are told that the speaker has an impairment "that causes social and linguistic problems." When receiving such a top–down instruction, listeners no longer interpret prenominal modifiers produced by the given talker as a meaningful cue to a contextual contrast (cf. Sedivy et al., 1999). With such a case of pragmatic impairment and a strictly rational model as two extreme ends of a continuum, talkers will often vary in terms of how much information they typically include in an utterance. Some talkers will be prone to provide additional descriptors while others' utterances are more succinct. Each talker, however, is likely to be relatively consistent. To the extent that these assumptions hold, flexibly adapting an expected form of an utterance for a given talker will prevent listeners from going astray when they encounter more or less information than what is *a priori* expected.

To test this hypothesis, we created an experimental paradigm with an Exposure Phase and a Generalization Phase. In the Exposure Phase the input from one of two speakers deviates from what is expected based on the rational speaker model. Specifically, that speaker does or does not use a scalar adjective (e.g., big/small) that would be necessary for singling out a referent, or if used, would provide redundant information (under- and over-modifying speakers given the rational model). We then examine in a Generalization Phase whether listeners derive different referential expectations for these two speakers (i.e., talker-specific expectation adaptation). In addition, we present a previously unseen set of adjectives in the Generalization Phase to examine the robustness of the adaptation process. We hypothesized that rational listeners would generalize from

[1] Our comments apply only to how people weight evidence that is presented verbally, as in the Linda problem. Many of the Tversky and Kahneman heuristics that lead to fallacies are based on the people's priors being distorted, e.g., overestimating the likelihood of events like airplane crashes, children being kidnapped, and infection from Ebola because the input, e.g., news coverage, is distorted. Thus people might make judgments and decisions that are objectively incorrect, even if their reasoning followed the principles of rational inference.

their experience, resulting in more accurate expectations for a wider range of linguistic expressions than those for which they have direct evidence. For example, upon observing utterances with referring expressions from a talker who provides over-specified expressions along one dimension (e.g., big/small), a listener might infer that that talker would also be more likely to over-specify along other dimensions (e.g., skinny/wide). (In Experiments 3 and 4, we provide a direct test of this prediction with adult native speakers of English.) We thus examine listeners' adaptation of referential expectations for uses of observed and unobserved adjectives.

One important factor that influences patterns of generalization is listeners' prior beliefs about how talkers might vary in their reference generation. For example, an instance of a seemingly over-specifying adjective can be compatible with at least two hypotheses: (1) the talker is incapable of making an optimally informative utterance (informativity-based generalization), or (2) the talker prefers to produce a longer utterance (length/form-based generalization). Also, listeners need to determine if the over-specification is confined to (1) the particular type of adjective, (2) adjectives in general, or (3) any form of modification. Moreover, one episode of sub-optimal language use could be indicative of the talker's overall pragmatic ability or it could be a random production error, in which case it would have little predictive power about future input. To avoid over- and under-generalization, rational listeners must evaluate the observed evidence against their prior beliefs to estimate how reliably it conveys information about the pragmatic competence of the talker (Xu and Tenenbaum, 2007; for a theoretical discussion on effects of prior beliefs in phoneme adaptation and generalization see Kleinschmidt and Jaeger, 2015). Based on this assumption, we predict a critical difference in how listeners generalize evidence of under- and over-specified utterances. Given the prevalent over-modification observed in natural discourse, a single instance of a redundant adjective use provides less reliable evidence that the speaker would be non-optimal in other domains of pragmatic language use compared to a single instance of under-specification. Therefore, we should see more conservative generalization (at the speaker-level) from evidence of over-specification compared to evidence of under-specification.

With the exception of pioneering work by Grodner and Sedivy (2011), talker- or context-specific adaptation and generalization of expectations have not thus far been studied extensively with respect to reference resolution (but see Arnold et al., 2007, for related discussion on comprehension of disfluencies). However, the importance of adaptation and generalization is increasingly appreciated in other domains of language processing. In particular, talker- and context-specific adaptation is crucial for comprehenders to navigate the problem of lack of invariance between the acoustic signal perceived and underlying linguistic categories such as phonemes. Some of this lack of invariance is due to random factors, such as errors in production and perception, but much is due to systematic factors, such as differences between speakers, dialects/accents, and speech conditions. A number of studies have demonstrated that listeners condition their perception of phonetic categories on talkers and their indexical features and learn to expect different acoustic features in the input for these different groups of talkers and different situations (e.g., Strand and Johnson, 1996; Niedzielski, 1999; Bradlow and Bent, 2008; Reinisch and Holt, 2014; for review see Drager, 2010; Kleinschmidt and Jaeger, 2015). Our framework shares a number of important properties with models developed to address the lack of invariance in speech perception. Most importantly, we view the problem of reference resolution as a form of systematic inference based on variable input in which listeners condition their inferences taking into account talker-specific information.

The remainder of the paper is structured as follows. We present results of four sets of experiments, in which we examine talker-specific generalization based on under-modified (Experiments 1 and 2), and over-modified (Experiments 3 and 4) utterances. We first establish that listeners will generalize information from a single pair of adjectives to unseen adjectives in a talker-specific manner based on observation of under-modified utterances (see Experiment 1: Talker-Specific Adaptation and Generalization Across Adjectives). We then tease apart two possible dimensions of talker-based generalization, which we call informativity-based and form-based generalization. A single observation of an under-modified utterance (e.g., "Click on the cup" in a presence of a big and a small cups) could be interpreted as evidence that the talker has a propensity to produce (1) under-informative expressions (i.e., informativity-based generalization) or (2) shorter expressions (form-based generalization). By introducing modified, yet under-informative utterances (e.g., "Click on the green cup" when the big and the small cups are both green), we demonstrate that whereas the generalization is primarily informativity-based some listeners more frequently made form-based generalizations (see Experiment 2A: Informativity-based vs. Form-based Generalization for Talker-Specific Adaptation). The preference for informativity-based generalization is magnified when the task is presented with an explicit instruction directing comprehenders' attention to differences between the talkers (see Experiment 2B: Effects of Adding a more Explicit Cue – Focus on Naturalness), suggesting that construal of the task influences how listeners generalize from the evidence that they observe.

We then turn to exposure to over-modified utterances. Given the prevalence of such utterances in simple referring tasks, we predict more conservative generalization across adjective types compared to cases with under-modified utterances. The results suggest that the over-modified utterances are indeed unlikely to trigger informativity-based generalizations (see Experiment 3: Talker-Specific Adaptation with Over-Informative Evidence) although comprehenders do register that the two talkers' utterances differ in length (see Ruling out an Alternative Explanation based on a Failure to Generalize overall for Over-Informative Utterances). This absence of informativity-based generalization persisted even when an extra manipulation highlighting the non-optimality of over-modifying utterances in referential communication was added (see Experiment 4: Drawing more Attention to the Fact that Over-Informative Information is not Helpful). In the General Discussion, we discuss an inference mechanism that provides a framework for explaining these different patterns of talker-specific generalizations of pragmatic information and suggests promising venues for future investigations.

EXPERIMENT 1: TALKER-SPECIFIC ADAPTATION AND GENERALIZATION ACROSS ADJECTIVES

We first asked whether listeners would generalize information from observed to unobserved (new) adjectives in a talker-specific manner. Importantly, because listeners are unlikely to be given explicit, top–down information about pragmatic competence under most circumstances, we wanted to determine whether they would generalize without being explicitly told that the talker was pragmatically impaired as they were in Grodner and Sedivy (2011). In the Exposure Phase we introduced listeners to two talkers and tasked them with selecting the unique referent of the talker's instruction from a set of four objects. The two talkers varied in their descriptions: only one talker used adjectives to pick out a unique referent (the modifying talker) while the other talker consistently used bare nouns (the non-modifying talker)[2]. In the Generalization Phase, we asked the listeners to guess which talker likely uttered transcribed instructions that were either modified (with new, or previously used adjectives) or unmodified. If listeners had generalized their assumptions about the talker's adjective use, they should attribute both the observed and new adjective use to the modifying talker, and the unmodified instructions to the non-modifying talker.

Methods

Participants

Thirty-two English-speaking adults residing in the USA were recruited online using the crowdsourcing platform Amazon Mechanical Turk (https://www.mturk.com/mturk/). Participants were compensated $1.00 for participating in the task[3, 4].

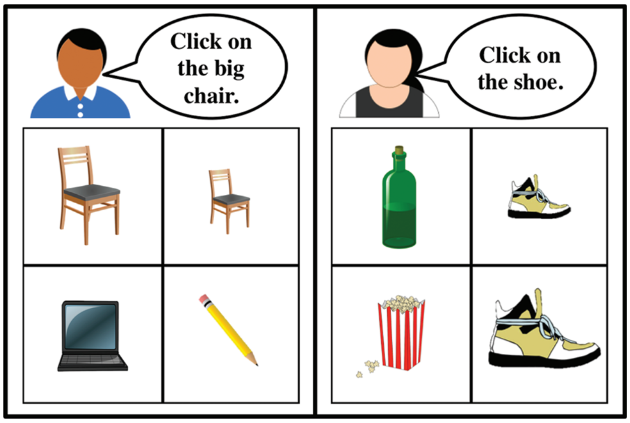


FIGURE 1 | An example of a trial (audio instructions are shown here in speech bubbles) with the modifying speaker (left) and the non-modifying speaker (right) from the Exposure Phase in Experiments 1 and 2. Participants clicked on the image in the grid to respond.

Materials

We created 44 two-by-two grids of images (20 for exposure items and 24 for test items). Each grid has a contrast pair of images differing from each other in one dimension and distinguished with a scalar adjective (e.g., a big cake vs. a small cake as in **Figure 1**). The other two images were singletons.

Two native speakers of American English (one male and one female) recorded 10 instructions each for the 20 exposure items. All the instructions were of the form "Click on the ____" and the two speakers recorded three versions for each item: a bare noun (e.g., "Click on the cake"), and with the adjectives *big* (e.g., "Click on the big cake") or *small* (e.g., "Click on the small cake"). 24 instructions were created for the Generalization Phase. One third of the modified instructions had the adjectives used in the Exposure Phase (four instructions each with *big* and *small*). The remaining two thirds of the modified instructions used new adjective pairs (four *tall*/*short*, four *skinny*/*wide*). Generalization instructions were presented as written scripts.

Procedure

In the *Exposure Phase,* participants were shown two-by-two grids of images. We provided a cover story that two naïve talkers had participated in a production task and produced instructions of the form "Click on the ___." The job of current participants was to follow these instructions and select one picture by clicking on it. On 10 of the trials one of the speakers (the modifying

[2] We chose to refer to the talkers modifying and non-modifying for the following two reasons. First, and most crucially, modifying/non-modifying and over-/under-informative are two orthogonal dimensions. A modified utterance, e.g., "Click on the big apple," can be over- or under-informative given a visual scene and comprehenders' construal of the task. (Similarly, in Experiment 2, a color-modified utterance is modified but under-informative.) In our experiments the two talkers are distinguished by the forms of the utterances they use. Participants could either attribute the difference between the talkers to differences in informativity or to preferences in utterance length. Therefore, it is more appropriate to use modifying and non-modifying to refer to the two talkers. Secondly, referring to the talkers as modifying/non-modifying provides a coherent thread across all of the experiments. With our current manipulation, the input sentences in the exposure phase are identical across experiment (i.e., one talker produces modification and the other does not). Whether these utterances are informative or not differs depending on the visual context (i.e., experiment).

[3] All of the experiments discussed in this paper were carried out in accordance with the recommendations of the Research Subjects Review Board at the University of Rochester, all participants gave consent to participate in the studies. They were given a digital copy of the consent form, and were asked to click a radio button labeled "I accept" to indicate that they had read and understood the consent form. All participants gave informed consent in accordance with the Declaration of Helsinki.

[4] Amazon's Mechanical Turk permits requesters to set selection criteria, and only the Turkers that meet the selection criteria are shown the task in their list of available HITs (Human Intelligence Tasks). For the purposes of this task we set the selection criteria to include a location restriction to users in the USA, who have an overall approval rate of 95% or higher, and who have not participated in any of the other experiments related to those described in this paper. Participants were told in the description of the task that we were looking only for native English speakers (having learned English at the age of five or younger). Participants were additionally excluded if they gave two or more incorrect responses on the modified Exposure Phase (EP) trials in Experiments 1 and 2, on either type of EP trial in Experiment 3, or non-modified EP trial in Experiment 4. Participants who viewed the HIT but did not complete the task were excluded from the analyses, and are not reported here.

talker) made a request using a prenominal adjective such as, "*Click on the big/small cake*" (five items with *big* and five items with *small*). On the remaining 10 trials the other talker (the non-modifying talker) produced instructions with bare nouns (e.g., "*Click on the cake*"). On each trial an avatar depicted which of the speakers the participant would hear on that trial (see **Figure 1** for an example of an Exposure Phase trial). The items were presented in a randomized order. The location of the target object, adjective (big vs. small), and gender of the modifying talker were counterbalanced across participants. Participants were instructed to make their best guess when they thought the speaker was unclear, or if they were uncertain. Participants were not given any feedback about their responses.

In the *Generalization Phase,* participants were told that they would read instructions that had been transcribed. Their task was to judge which of the two speakers was more likely to have produced the instruction and click on the corresponding avatar (**Figure 2**). 12 of the 24 instructions contained a modifying adjective. Four of the modified instructions contained the same adjectives as in the Exposure Phase (big/small); eight contained new scalar adjective pairs (two skinny/two wide, two tall/two short). On the remaining 12 trials the instructions were unmodified. These items were presented in a randomized order. The adjective-object pairing and type of instruction (modified vs. unmodified) was counterbalanced across participants. After making their selection participants were asked to rate how confident they were in their response on a five-point scale (1 = not at all confident, and 5 = completely confident).

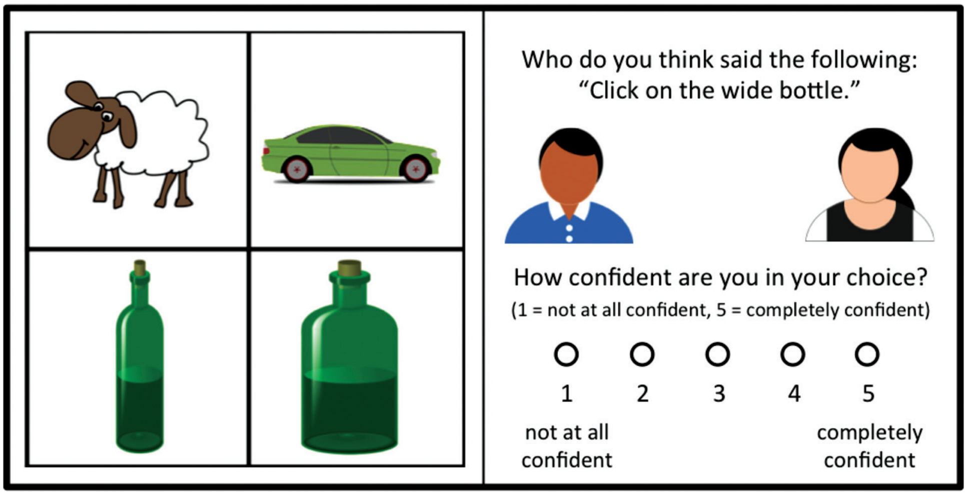
Results and Discussion

Choices in the Generalization Phase are plotted in **Figure 3**. Participants selected the modifying-speaker, who used *big/small* in the Exposure Phase, for the sentences with *big/small* (83%), and the non-modifying speaker in the unmodified trials (80%). Choice patterns for new adjectives were almost identical to those for exposure adjectives: 84 and 84% for skinny/wide, and 83 and 84% for tall/short. We constructed a mixed-model logistic regression of the responses given for the modifying speaker in the Generalization Phase with Adjective (exposed or new), and Instruction Type (modified or non-modified) as the fixed effects, and subject and item as the random effects[5]. We based our model on the recommendations for maximal Linear Mixed Effects Model (LMEM) as suggested by Barr et al. (2013) which takes into consideration the maximal random effects structure by including by-subject (Adjective and Instruction Type) and by-item (Instruction Type) random intercepts and slopes. We used the glmer function in lme4 in R, and specified a BOBYQA optimizer (Bates et al., 2015). As predicted, Instruction Type was the only significant predictor of whether participants would choose the modifying speaker ($\beta = 5.84$, $p < 0.001$). There were no reliable predictors of the confidence ratings (*ps* = 1), indicating that participants were equally certain (modified mean = 3.8/5; non-modified mean = 3.81/5) about their responses regardless of the Instruction Type and Adjective (exposed or new).[6]

The results support two predicted effects of the exposure items. First, participants reliably track talker-specific usage patterns of adjectives and choose the modifying talker for new instructions with previously observed adjectives (i.e., big/small). Second, participants generalized their assumptions to previously unobserved scalar adjectives and chose the modifying talker for instructions with new scalar adjectives. Thus listeners

[5] The same factors were used in a mixed-effects logistical regression for all experiments unless noted otherwise.

[6] In this and all of the other experiments we report below, the main effects (Instruction Type and Adjective) were not a significant predictor of the confidence rating scores. Therefore, we do not discuss these results in subsequent experiments; instead we report the data in the Figures.



Who do you think said the following:
"Click on the wide bottle."

How confident are you in your choice?
(1 = not at all confident, 5 = completely confident)

○ 1 not at all confident
○ 2
○ 3
○ 4
○ 5 completely confident

FIGURE 2 | In the Generalization Phase participants saw 2x2 image grids (left) above the transcribed instructions, avatars that represented the two speakers, and a confidence rating scale (right).

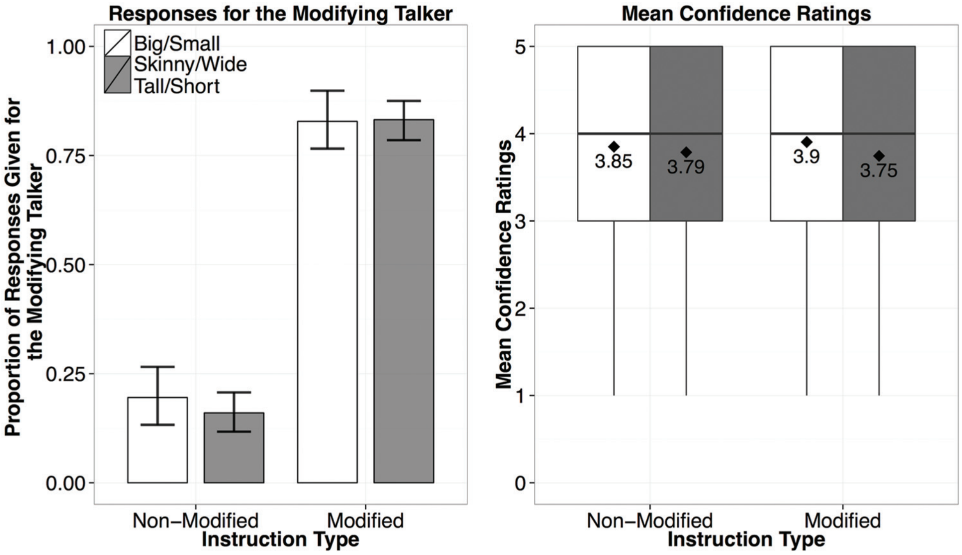


FIGURE 3 | Results from Experiment 1, showing the proportion of responses given for the modifying talker by Instruction Type (left), and the confidence ratings for responses by Instruction Type (right; diamond dots reflect the mean rating out of 5).

quickly adapt their expectations for a particular talker's referring expressions.

However, these results are compatible with at least two classes of accounts. Participants could have inferred that one speaker provided the sufficient amount of information to uniquely refer, while the other did not (Informativity-based generalization). Alternatively, participants could have inferred that one of the speakers was more likely to produce modified utterances (Form-based generalization). In Experiment 2, we modified the instructions in the Generalization Phase to investigate which account better predicts listeners' adaptation/generalization behavior.

EXPERIMENT 2: GENERALIZATION FROM UNDER-INFORMATIVE EVIDENCE

Experiment 2A: Informativity-Based vs. Form-Based Generalization for Talker-Specific Adaptation

Experiment 2A examined whether participants inferred that one of the speakers was more or less informative (Informativity-based generalization) or generalized based on utterance length (Form-based generalization). We replaced the bare noun instructions in the Generalization Phase of Experiment 1 with orthogonal color adjectives (e.g., *Click on the green car* when both cars in the scene are green). If generalization is based on informativity, participants should select the previously non-modifying (under-informative) speaker. If, however, generalizations are form-based (i.e., based solely on whether or not a speaker had used an adjective), participants should select the modifying speaker on both the color-adjective trials and the scalar adjective trials.

Methods

Participants

Thirty-three English-speaking adults residing in the USA who had not previously participated in a study in this series were compensated $1.00 for taking part in the task on Amazon Mechanical Turk. We applied the same exclusion criteria as what we used in Experiment 1.

Materials

The visual and the audio materials for the Exposure Phase were identical to those used in Experiment 1. We constructed 12 new instructions for the Generalization Phase by replacing the non-modified instructions with instructions containing color adjectives. These instructions were paired with two-by-two grids with the contrastive item pair that differed in size along the same dimensions as the scalar adjectives used in the scalar modified trials, but did not differ in color. Thus, these color-modified instructions such as "*Click on the green bottle*" would not pick out a unique referent. The remaining 12 scalar-modified instructions such as "*Click on the wide bottle*" (**Figure 2**), carried over from Experiment 1, would pick out a unique referent. Thus, all instructions in the *Generalization Phase* contained either a scalar or a color adjective. Experiment 2A used the same instructions as Experiment 1.

Procedure

Procedure was identical to Experiment 1. Participants were not given feedback on their responses and asked to rate confidence in their selection after each item in the Generalization Phase.

Results and Discussion

Participants' responses were similar to those in Experiment 1 (see **Figure 4**). For both observed and new scalar adjective types,

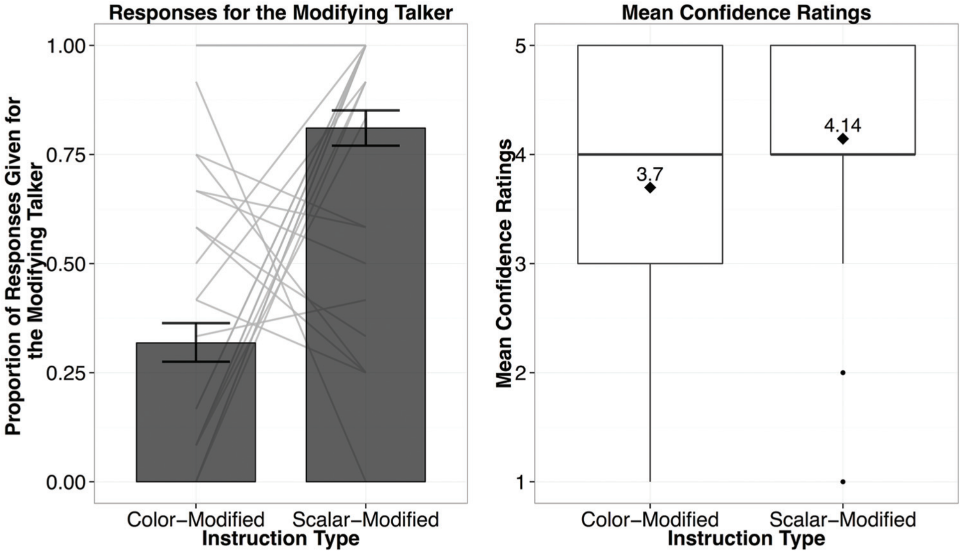


FIGURE 4 | Results from Experiment 2A, showing the proportion of responses given for the modifying talker by Instruction Type (left; light gray bars reflect individual participant means), and the confidence ratings for responses by Instruction Type (right; diamond dots reflect the mean rating out of 5).

they primarily picked the modifying speaker (81%). However, on the color-modified trials, participants showed preferences for the non-modifying talker (68%). These results show that participants are making informativity-based generalizations, choosing the previously non-modifying talker for modified yet under-informative instructions. The mixed-effects logistic regression found that the only reliable predictor of whether the listener chose the modifying speaker on a given trial was Instruction Type [scalar-modified (informative) vs. color-modified (under-informative); $\beta = 6.428$, $p < 0.001$].

In sum, these results suggest that not only have participants discovered that there is something linguistically different between the talkers, but also that one of these talkers was using pre-nominal modification to provide information that allows unique reference, whereas the other was not. Participants were willing to attribute new color-modified utterances to a talker they have never heard using color adjectives to modify. Thus participants have inferred that only one of the talkers uses modification to provide sufficient information for unique reference.

Comparison of Experiments 1 and 2A

We compared Experiments 1 and 2A using a mixed-effects logistic regression analysis with Experiment (1 vs. 2A), Adjective (exposed vs. new), and Instruction Type (under- vs. concisely-informative) as fixed effects, and subject and item as random effects, including by-subject (Adjective and Instruction Type) and by-item (Experiment and Instruction Type) random intercepts and slopes. We used a model with the correlations between the random slopes and random intercepts removed as recommended by Barr et al. (2013) when the maximal model fails to converge. We found evidence that is comparable to what we found in Experiment 1 for talker-specificity: Instruction Type was a predictor of the responses for the modifying talker, $\beta = 5.752$, $p < 0.001$. In addition, there was a predictive effect of Experiment ($\beta = 0.807$, $p = 0.05$). The predictive main effect of Experiment is likely driven by a smaller percentage of responses for the non-modifying speaker on the color-modified trials in Experiment 2A in comparison to the percentage of responses for the non-modifying speaker on the non-modified trials in Experiment 1.

To take a closer look at patterns of responses by individual participants, we plotted mean proportion of choice of the modifying talker with the light gray connecting bars in **Figure 4**. It is evident that there is a substantial amount of individual variation: while the majority of participants responded in a way that reflects informativity-based generalizations (lower responses for the modifying speaker on color-modified trials, and higher responses on the scalar-modified trials), some participants seem to be making form-based generalizations, as noted by relatively invariant responses for the modifying speaker across both trial types. We suspected that the individual differences stemmed from the fact that participants differed from each other in terms of their construal of the current task. Some might have assumed the goal of the task was to evaluate the clarity and helpfulness of the instructions, which would have encouraged them to focus on informativity of the instructions. Others might have tried to match instructions in terms of their formal similarity. To test this idea, we made this assumption explicit in our instructions to see whether that would affect patterns of participants' responses.

Experiment 2B: Effects of Adding a more Explicit Cue – Focus on Naturalness

In Experiment 2B, we repeated Experiment 2A but added an extra instruction that asked the listeners to pay close attention to potential speaker differences in "clarity" and "naturalness." We hypothesized that the explicit instruction would increase informativity-based generalizations by highlighting the fact that instructions vary along the dimension of helpfulness in picking out a unique referent. We also report two follow-up analyses. First, we present a mixture model analysis of a combined dataset from Experiments 2A and 2B to further investigate the effect of explicit instructions. We then present data from a follow-up experiment in which comprehenders observed informative uses of color-adjectives and under-informative uses of scalar adjectives in the Generalization Phase. We predict a similar – possibly slightly diminished – degree of informativity-based generalization, which would rule out the possibility that the generalization is limited to a particular adjective type.

Methods

Participants

Thirty-two English-speaking adults residing in the USA who had not previously participated in a study in this series were compensated $1.00 for taking part in the task on Amazon Mechanical Turk. We applied the same exclusion criteria as we used in the previous experiments.

Materials

Identical to Experiment 2A.

Procedure

The procedure was identical to Experiment 2A, except that participants received audio instructions instead of the written instructions used in Experiments 1 and 2A. This was to ensure that they heard all of the details of the instructions. Participants were told that the goal of the task is to select instructions by speakers that made the clearest or most natural instructions, and that at the end of the task there would be an opportunity for them to provide feedback on whether either of the speakers' instructions were unusual in any way. At the end of the experiment participants were asked to indicate which speaker they thought was the clearest and most natural sounding, and then were asked to describe why they thought the other speaker was less clear or natural.[7]

Results and Discussion

When attention was called to the clarity of the two speaker's utterances, participants' responses showed more pronounced trends toward informativity-based generalizations. Participants selected the talker who previously did not use adjectives in the Exposure Phase in the (under-informative) color-modified trials, 88% of the time. As in Experiment 2A, the only reliable predictor of whether the modifying talker was chosen was Instruction Type ($\beta = 8.109$, $p < 0.001$), meaning that yet again we see evidence for participants overall generalizing based on informativity.

Comparison of Experiments 2A and 2B

We compared Experiments 2A and 2B using a mixed-effects logistic regression analysis with Experiment (2A vs. 2B), Adjective (exposed vs. new), and Instruction Type (scalar- vs. color-modified) as fixed effects, and subject and item as random effects, including by-subject (Adjective and Instruction Type) and by-item (Experiment and Instruction Type) random intercepts and slopes. We used a random-slopes only model as suggested by Barr et al. (2013), rather than a maximal model, or a model with independent random slopes and intercepts, as they both failed to converge. Instruction Type ($\beta = 5.76$, $p < 0.001$), the interaction of Experiment by Instruction Type ($\beta = 2.888$, $p = 0.05$), and a three-way interaction between the fixed effects ($\beta = 2.366$, $p = 0.01$) were significant predictors of whether or not participants chose the modifying speaker. The interaction of Instruction Type and Experiment supports the idea that explicit instructions biased participants to generalize more on informativity: with explicit instructions fewer participants attributed the color-modified instructions to the non-modifying (under-informative) talker, and fewer participants attributed the scalar-modified instructions to the modifying talker (see **Figure 5**). The explanation for the interactions will become clear in the following analyses.

Mixture Model Analysis of Experiments 2A and 2B

Because we hypothesized that explicit instructions would result in more informativity-based generalizations, we tested for patterns of generalization across participants. We did so by fitting multivariate mixture models to the data. Separate models were fit for each Instruction Type in each of the conditions for 1-6 components using the mixtools package (Benaglia et al., 2009) in R, which uses expectation maximization (EM) to estimate the optimal parameter values.

On the scalar-modified trials, participants primarily attributed these instructions to the modifying talker, and even more so in Experiment 2B. In Experiment 2A our mixture model analysis found that the majority (73%) of the participants selected the modifying talker for the scalar adjective trials on average 98.2% of the time, and the remaining 27% of participants selected the modifying talker on average 35% of the time. In Experiment 2B the model found that the majority (88%) of the participants selected the modifying talker for the scalar adjectives on average 98.2% of the time. The remaining 12% of the participants selected the modifying talker on average 59% of the time.

For color-modified trials, in Experiment 2A, a three-component model fit the data significantly better than the either the one-component [$\chi^2(6) = 373.2$, $p < 0.001$] or the two-component [$\chi^2(3) = 18.8$, $p < 0.001$] models. Participants responses fell into the following three categories: (1) 12% of the participants selected the modifying talker for these trials 98% of the time (evidence for form-based generalizations); (2) 30% of participants selecting the modifying talker 57% of the time (approximately chance-like behavior, indicating that they thought either speaker could have produced these instructions with equal likelihood); and (3) the remaining 58% of the participants picked the modifying talker only 5% of the time (evidence for informativity-based generalizations). In contrast, in

[7] The majority of the respondents indicated that that they thought that the less clear speaker did not provide necessary adjectives / enough information to help pick out objects, only four of the 32 respondents indicated that they thought one of the speakers was less clear or natural for other reasons, such as the quality of the speaker's voice ($n = 2$), or thought they were both quite clear ($n = 2$). We take this as evidence that the participants overall as a group interpreted the instructions to be about the informativity of the instructions rather than features of the speakers' voices.

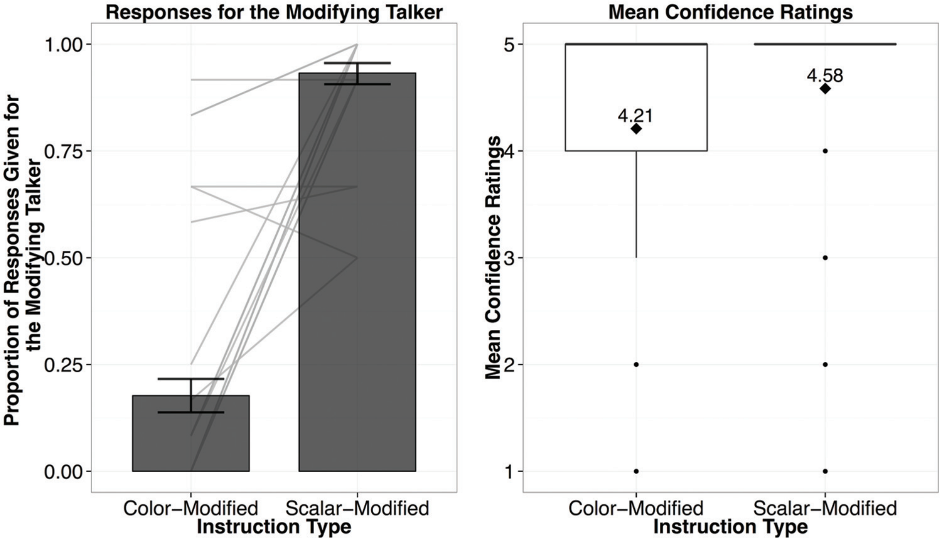


FIGURE 5 | Results from Experiment 2B, showing the proportion of responses given for the modifying talker by Instruction Type (left; light gray bars reflect individual participant means), and the confidence ratings for responses by Instruction Type (right; diamond dots reflect the mean rating out of 5).

Experiment 2B a two-component model fit the data significantly better than a one-component model [$\chi^2(3) = 305.1$, $p < 0.001$] or a three-component [$\chi^2(3) = 0.14$, $p = 1$] models. Individual participants responded in one of two ways: (1) 81.2% of the participants picked the modifying speaker only 5% of the time (evidence for informativity-based generalizations); and (2) the other 18.8% of the participants selected the modifying speaker for these trials 75% of the time (evidence for form-based generalizations).

This analysis reveals that there is more variability in participant response patterns in Experiment 2A, compared to Experiment 2B. This can be seen in the tighter clustering pattern toward the top left corner in Experiment 2B in **Figure 6**. If listeners generalizations are informativity-based, we expect results to cluster in the top left (meaning that an individual always picked the modifying speaker for the scalar-modified trials, thus approaching 1, and almost never for the color-modified trials, approaching 0), whereas if they were form-based we expect clustering in the top right corner (where the proportion of responses for the modifying speaker is near 1 for both instruction types). In sum, calling attention to the quality of the instructions made listeners more willing to infer that the non-modifying speaker would be less pragmatically optimal and therefore *more* likely to use an under-informative color-adjective.

Ruling out an Alternative Explanation based on Adjective Class

One possible concern is that the results in Experiments 2A and 2B could be due to listeners' tendency to associate a particular talker with a particular adjective type. Participants might have assumed that one of the talkers liked to use scalar adjectives and the other non-scalar adjectives. While associations like these are not attested in any previous research, it is possible that participants in the current study might have inferred that, at the very least, one of the speakers was more likely to use scalar adjectives than the other. To rule out this possibility, we conducted an additional version of Experiment 2 ($n = 32$) in which the items in the Generalization Phase contrasted in color rather than a scalar dimension. That is, in contrast to Experiments 2A and 2B, color-adjectives in the Generalization Phase were helpful in selecting a unique referent whereas scalar adjectives were not. If participants are generalizing based on informativity they should attribute the contrastive, color-modified instructions to the modifying speaker.

As in Experiments 2A and 2B, we found that the only reliable predictor of whether participants selected the modifying speaker was Instruction Type ($\beta = 6.459$, $p = 0.05$). Participants selected the modifying speaker for the color-modified informative instructions 84% of the time. Participants also selected the modifying speaker for the non-informative scalar-modified instructions. However, as we predicted, they did so less than for the color-modified instruction (58% compared to 84%).

What accounts for the relatively high selection rate (58%) of the previously modifying (informative) talker for the under-informative scalar-modified instructions? We have two hypotheses. First, listeners have weighted heavily their direct observation of one talker using scalar adjectives in the Exposure Phase. This might have made it difficult for them to inhibit the expectation that the previously modifying talker would continue to use scalar adjectives. Second, unlike color-modifiers, talkers in general rarely produce scalar-modifiers in non-contrastive situations (Pechmann, 1989; Belke and Meyer, 2002; Sedivy, 2003; Brown-Schmidt and Tanenhaus, 2008; Viethen et al., 2012). Therefore, a listener would not expect a speaker who did not use adjectives at all to begin producing a non-contrastive

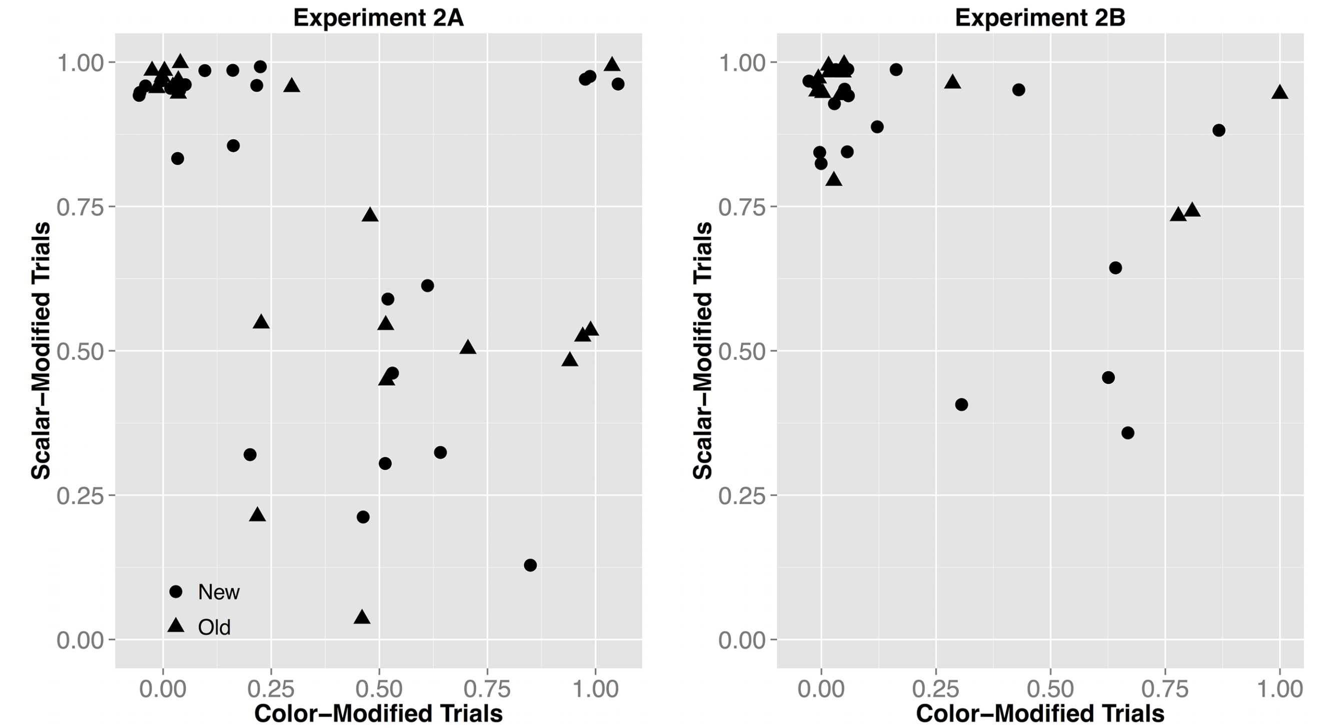


FIGURE 6 | Proportion of responses given for the modifying talker for color- by scalar-modified trials for each individual subject in Experiment 2A and Experiment 2B. Informativity-based generalizations (dots expected to pattern in the top left) are observed when listeners primarily select the modifying talker for the scalar-modified trials (proportion approaching 1.0) and rarely for the color-modified trials (proportion approaching 0.0). Form-based generalizations (dots expected to pattern in the top right) are observed when listeners primarily select the modifying talker for both trials (proportions approaching 1.0).

scalar-modified utterance compared to a non-contrastive color-modified utterance. In sum, these results provide additional support for our claim that participants were paying attention to the informativity of the talkers. As the same time, participants may have different expectations for different classes of adjectives (e.g., scalar- vs. color-adjectives) in terms of how reliably they would support an informativity-based generalization.

EXPERIMENT 3: TALKER-SPECIFIC ADAPTATION WITH OVER-INFORMATIVE EVIDENCE

As we noted earlier, speakers rarely under-modify (except in highly collaborative tasks; Brown-Schmidt and Tanenhaus, 2008). Do listeners' prior beliefs based on general statistics like these have any influence on ways in which they adapt their referential expectations? If so, how? As we mentioned in the introduction, the prevalent over-modification observed in natural discourses in contexts like the ones we used should lead to more conservative generalization compared to cases of under-modification. In particular, listeners might consider that a single instance of a redundant adjective use is not a good predictor of the same speaker's future pragmatic language use.

Integration of prior likelihoods into statistical inferences has proven effective in generalizing information meaningfully based on a limited amount of the input. For instance, word learners generalize information about novel word-referent mappings (e.g., "blicket" for a novel object) based on their prior beliefs about how nouns are used, who provided the data, and how the evidence is sampled (Xu and Tenenbaum, 2007). They then make inferences about how readily an observed word-referent mapping should be generalizable to other referents of the same kind rather than being restricted to a unique individual or property. Thus, by integrating relevant prior beliefs, listeners are able to evaluate the input with respect to how reliably it can predict previously unseen data, which helps reduce the chance of over- or under-generalization due to over-fitting their expectations to data observed locally.

Methods

Participants

Thirty-two English-speaking adults residing in the USA who had not previously participated in a study in this series were compensated $1.00 for taking part in the task on Amazon Mechanical Turk. We applied the same exclusion criteria as what we used in previous experiments. An additional participant completed the task but was excluded from the analysis for giving incorrect responses during the Exposure Phase.

Materials

Audio stimuli for the Exposure Phase (20 items) were identical to those in Experiments 1 and 2. Visual stimuli were modified

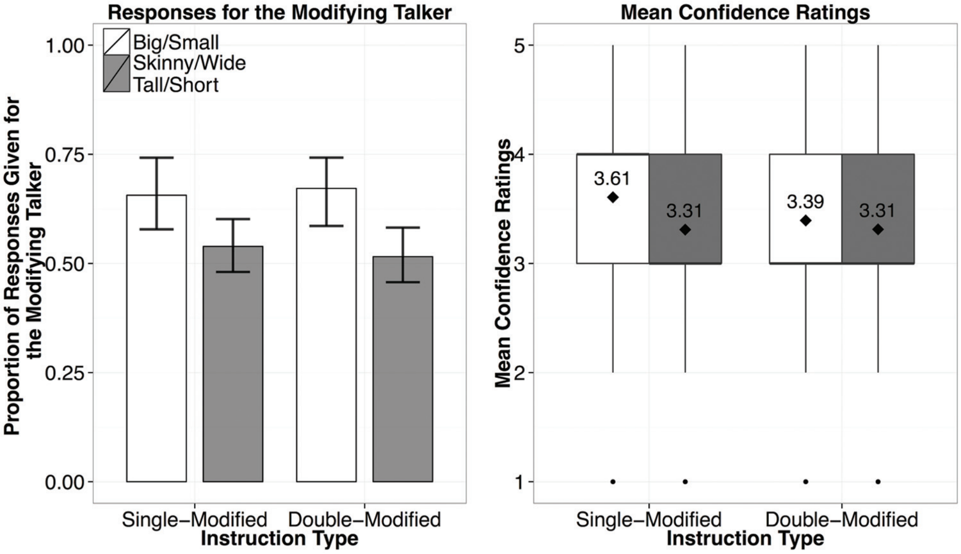


FIGURE 7 | Results from Experiment 3, showing the proportion of responses given for the modifying speaker by Instruction Type (left), and the confidence ratings for responses by Instruction Type (right; diamond dots reflect the mean rating out of 5).

so that there was no size contrast pair and each two-by-two grid consisted of four singleton images. This manipulation rendered a non-modifying instruction to be concisely informative (e.g., “Click on the cake”), and a modified instruction to be over-informative (e.g., “Click on the big cake”). Visual stimuli in the Generalization Phase (24 items) were identical to those in Experiments 1 and 2, containing a visual contrast pair. 12 of the 24 items were associated with a single-modified instruction (e.g., “Click on the wide bottle”) and the rest were associated with a double-modified instruction with both an informative scalar adjective and a redundant color adjective (e.g., “Click on the wide green bottle”). We predicted that if listeners were generalizing on informativity, rather than form, that they should attribute the single-modified (concisely informative) instructions to the previously non-modifying speaker, and the double-modified instructions (over-informative) to the modifying speaker, who appears to be habitually over-informative.

Procedure

As in the previous experiments, participants completed all the 20 exposure trials and 24 generalization trials consecutively. Participants read the same instructions used in Experiments 1 and 2A. Participants rated their confidence on each trial.

Results and Discussion

In contrast to the cases with an under-modifying talker (i.e., Experiments 1 and 2), the results from Experiment 3 show no clear evidence for informativity-based generalization (**Figure 7**). In a mixed-effects logistic regression Instruction Type (concise- or over-modification) was not a reliable predictor ($p > 0.1$), nor was the interaction of Instruction Type and Adjective ($p > 0.1$). The only reliable predictor of when participants would choose the modifying (over-informative) speaker was whether a previously encountered or a new scalar adjective was used ($\beta = 0.959$, $p = 0.05$): participants were more likely to choose the modifying speaker if a previously exposed adjective (*big* or *small*) was used (67%) regardless of whether it was used with a color adjective. For the new adjectives, there was no clear trend for participants to attribute the use of the new adjectives to either speaker, attributing them equally to both speakers (choosing the modifying speaker in for 52% of the responses).

As predicted, we found an asymmetry between the cases of under-modification and over-modification, in which listeners do not seem to make talker-specific informativity-based generalizations from exposure to over-modified instructions. This null effect with over-informative input was, however, somewhat surprising given the reliable effects of talker informativity found in Experiments 1 and 2. Before we conclude that this pattern of results is due to participants being more conservative about generalizing from over-informative utterances, we need to rule out another possibility. Perhaps participants did not notice that one of the talkers was over-modifying in the Exposure Phase.

Ruling out an Alternative Explanation based on a Failure to Generalize overall for Over-Informative Utterances

Unlike the under-modifying instructions used in Experiments 1 and 2, over-modifying instructions do not create referential ambiguity. Thus, if talker-specific adaptation requires an observation of a clear “error” signal based on possible miscommunication, then the manipulation we used might have been too subtle to trigger adaptation, To address this possibility we conducted a follow-up experiment ($n = 32$), modeled on Experiment 1 to see if we could observe form-based generalizations. Participants observed the same two speakers describing images from a two-by-two grid that was comprised entirely of unrelated singleton images (as in Experiment 3) in both the Exposure Phase and the Generalization Phase. Thus, in both the Exposure Phase and the Generalization Phase modified instructions were over-informative, and non-modified instructions were concisely informative. The results demonstrated that participants were more willing to attribute the over-modified utterances to the modifying speaker (85% compared to 18% for the non-modified), regardless of the adjective used ($\beta = 10.27$, $p < 0.001$). This makes it unlikely that participants in Experiment 3 were simply not aware of talker-differences in the Exposure Phase. It is, however, still possible that they did not regard over-informative utterances to be communicatively sub-optimal because they did not cause any referential ambiguity.

To see whether this might be the case for our instruction, we looked at responses in the follow-up questionnaire in Experiment 3 (identical to that of Experiment 2B), which asked participants to comments on the clarity and naturalness of the two talkers' instructions. Participants were divided as to which talker they preferred: some participants found the over-modifying talker to be clearer and more helpful (23%); others considered the over-modifying talker to be redundant and potentially confusing (45%). The remaining participants commented on the quality of the speakers' voices, the recordings, or gave no response (32%). Thus the asymmetrical treatment of under- and over-modifying utterances could be due in part to listeners not considering over-modifying utterances to be communicatively sub-optimal. This would make it less likely for them to expect similar behavior from the same talker across different contexts. In Experiment 4, we manipulated the Exposure phase of Experiment 3 to highlight the fact that producing over-modifying instructions can, at least in some cases, hinder referential communication.

EXPERIMENT 4: DRAWING MORE ATTENTION TO THE FACT THAT OVER-INFORMATIVE INFORMATION IS NOT HELPFUL

In Experiment 3, and in previous research (Engelhardt et al., 2006; Arts et al., 2011), listeners have been shown to treat some instances of over-specification as facilitatory. In Experiment 4, we introduced two modifications to the paradigm used in Experiment 3, with the intention of highlighting the potential pitfalls of over-modification in the current referential task. First, we truncated 50% of the audio instructions such that the concise referential expressions communicated sufficient information for unique referent identification (e.g., "Click on the ca-" when a target is "camera") whereas the over-modified expressions do not (e.g., *"Click on the sma-"* when there is more than one small referent in a visual scene). Second, after each trial, we provided feedback identifying the talker's intended referent. We implemented these changes to emphasize the fact that producing a superfluous adjective can result in referential ambiguity.

Methods

Participants

Thirty-four English-speaking adults residing in the USA who had not previously participated in a study in this series were compensated $1.00 for taking part in the task on Amazon Mechanical Turk. We applied the same exclusion criteria as what we used in the previous experiments.

Materials

Visual and audio stimuli were identical to Experiment 3 except for the following three changes in the Exposure Phase. First, the audio instructions were truncated mid way: five out of the 10 unmodified instructions were cut off after the onset syllable of the noun (e.g., "Click on the ca-" when a target is "camera"), the remainder were truncated mid-word after the second consonant (e.g., *"Click on the cam-"*). Five of the 10 modified instructions were truncated after the adjective (e.g., *"Click on the small"* when there is more than one small referent in a visual scene), and the remaining modified instructions were truncated after the onset syllable of the noun (e.g., *"Click on the small ca-"*)[8].

Secondly, half of the Exposure Phase trials contained two-by-two grids with a contrast pair, and half contained four singleton items. Crucially the instructions produced by both speakers never referred to an item from the contrasting pair. Third, after each trial participants were shown which item the speaker was originally asked to describe. On the trials where the recording was cut off after the adjective, it was expected that the use of the redundant scalar adjective would be seen as misleading. An example trial from Experiment 4 can be seen in **Figure 8**. The Generalization Phase was identical to that of Experiment 3.

Procedure

The procedure was the same as in Experiment 3.

Results and Discussion

Despite the changes we made to the Exposure Phase, the results were nearly identical to those in Experiment 3 (**Figure 9**). In a mixed-effects logistic regression Instruction Type was not a significant predictor of whether participants chose the modifying speaker ($\beta = 0.169$, $p > 0.1$), However, whether an exposed or a new adjective was used was a significant predictor ($\beta = 0.635$, $p = 0.05$). Participants were overall more likely to attribute the instructions containing the words *big* or *small* to the modifying

[8] We chose this manipulation over more overt ways of highlighting over-informativity of the modifying-talker's utterances such as saying "This speaker has only four words to spend" for the following two reasons. First, we were concerned that a word limit ("This speaker has up to four words to spend") would make the task highly unnatural. We could not find a naturalistic context in which a speaker needs to control the number of words in a spoken sentence and we did not know how our participants would construe a situation like that. Second, if we introduced a word or a time limit in the introduction, listeners would likely expect the speaker to alter the syntactic structure of their instruction (e.g., "big apple" rather than "Click on a big apple") and/or increase their speech rate. However, including modulations like these would introduce other unknown factors such as (un)intelligibility of instructions, which would make the experiment less comparable to the other experiments reported in the current paper. For these reasons, we used the cover-story, which supports the assumption that the speakers were unaware of this problem and hence their syntactic and phonological formulations of instructions were consistent with those in other experiments.

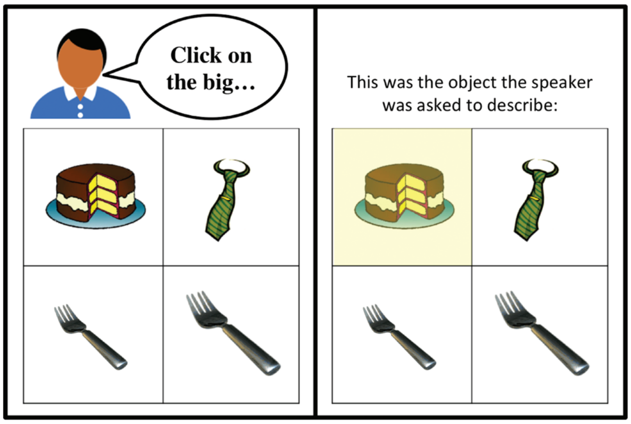


FIGURE 8 | An example of a trial (audio instructions are shown here in speech bubbles) with the modifying speaker (left) in the Exposure Phase in Experiment 4. Participants clicked on the image in the grid to respond. After each trial participants were shown which of the four images the speaker was asked to describe **(right)**.

speaker (62%) than the instructions containing new adjectives (50%), regardless of the Instruction Type (single modified, or double-modified, as noted by the lack of an interaction predictor). The results of Experiment 4 strongly suggest that the absence of evidence for generalization in a talker-specific manner from over-informative evidence is not due to listeners failing to register its sub-optimality in the communicative task at hand. It is more likely that listeners did not weigh evidence of over-informative instructions as much as that of under-informative ones, leading to more conservative generalization of their referential expectations at the talker level.

GENERAL DISCUSSION

We proposed that listeners adjust their expectations for how individual talkers might vary in their uses of referring expressions. This permits listeners to maintain the assumption that talkers are generally rational, rather than only sometimes rational, while allowing them to flexibly cope with the variability in speakers' referring expressions. We presented four sets of talker-selection experiments, examining if, and, if so, how, listeners adapt and generalize their referential expectations according to the observed input. We examined cases in which one of the two talkers produced either an under-modified or an over-modified utterance for a referent in a visually co-present context.

Summary of Results and Contribution

Under-Modification

With under-modified instructions, we found clear evidence that listeners adapted to talker-specific differences in the use of pre-nominal adjectives along the dimension of informativity. When a talker under-modified, listeners inferred that that talker would not formulate an informative utterance with other scalar adjectives (Experiments 1 and 2). Moreover, they considered the possibility that the same talker would formulate under-informative utterances with color adjectives (Experiment 2). This demonstrates that listeners generalized the information given in the Exposure Phase based on informativity. Importantly, we found evidence for informativity-based generalizations even when the evidence was implicit. However, the proportion of informativity-based generalizations increased when the instructions directed participants to pay attention to the clarity of the talkers' instructions.

Our results with under-modified instructions build on Grodner and Sedivy's (2011) results in three ways. First, Grodner and Sedivy reported that talker-specific modulation of pragmatic processing, in their paradigm, required explicit top–down information about the speaker being pragmatically impaired (though they mention that there was still a trend when top–down information was not provided). In contrast, our experiments with under-modification induced talker-specific adaptation of referential expectation without such an explicit instruction. This suggests that listeners are in principle capable of modulating their expectations based on bottom–up input alone. Second, our results contrasted generalizations based on under- and over-informative utterances, and thereby shed light on the dimensions over which listeners are generalization. Third, we show that, depending on the participant and the task, generalization can be more or less form-based and informativity-based.

Over-Modification

With a speaker who regularly over-modifies, we did not observe informativity-based generalization (Experiment 3). Even when a superfluous modifier was clearly unhelpful in reference resolution, listeners did not assume that overly informative utterances were a characteristic of an individual talker (Experiment 4). Listeners did make talker-specific generalizations, but they were overwhelmingly form-based rather than informativity-based. Our results with over-modified instructions might seem superficially inconsistent with Grodner and Sedivy's evidence for talker-specific adaptation with over-modified instructions. Recall, however, that some researchers (Arnold et al., 2007; Grodner and Sedivy, 2011) noted that they obtained robust results only when they explicitly called attention to the speaker's overall linguistic incompetence. Thus our results can be viewed, akin to the findings of Grodner and Sedivy, as supporting the suggestion that generalization from over-modification is strongest when there is top–down information that establishes a causal link between the redundant use of a prenominal modifier and the pragmatic propensities, or even, the linguistic competence, of the talker.

The asymmetry between the results with under-modification and over-modification is particularly striking. It provides strong support for the assumption that generalization takes into account prior beliefs based on the statistical structure of the data, in this case typical patterns of modification. We discussed the possibility that informativity-based generalizations might be weaker with over-modification than with under-modification because under-modifying utterances interfere more with communication in the task at hand. Admittedly, under-modifying utterances do

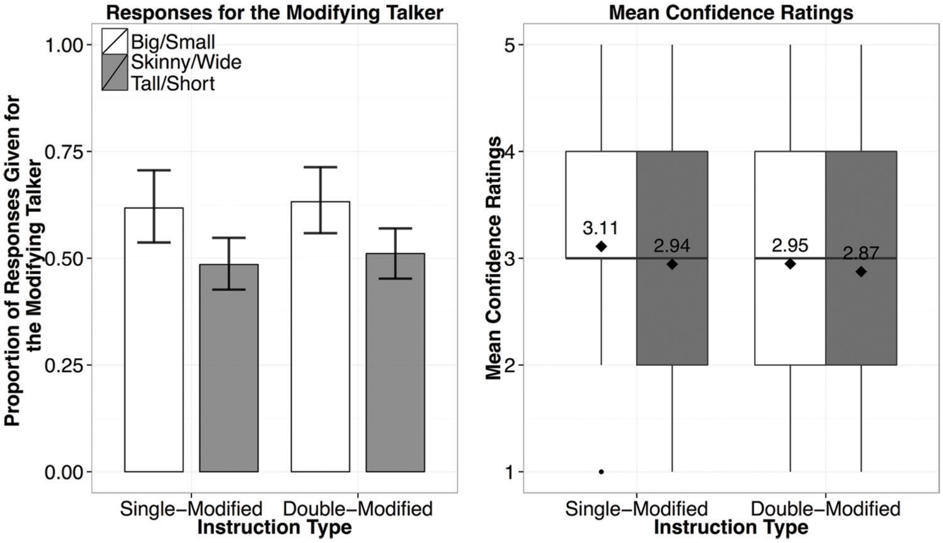


FIGURE 9 | Results from Experiment 4, showing the proportion of responses given for the modifying speaker by Instruction Type (left), and the confidence ratings for responses by Instruction Type (right; diamond dots reflect the mean rating out of 5).

not allow listeners to single out a unique referent, which calls attention to the sub-optimality of those utterances. In contrast, over-modifying utterances, allow listeners to pick out an intended referent. In fact, providing additional information is often considered a sign of helpfulness (Engelhardt et al., 2006; Arts et al., 2011). The likelihood of communicative error in a given context cannot, however, account for our pattern of results. The truncated utterances in Experiment 4 created referential ambiguity, drawing attention to the fact that by including superfluous material the over-modifying talker generated referring expressions that resulted in communication failure.

It is possible that general inferences about informativity from over-modification might emerge only with more robust manipulations in highly collaborative tasks, e.g., in a video game task where timely actions based on communication with a partner are required. Alternatively, because offline measures do not capture real-time expectations, an online measure might reveal effects that are not captured in offline measures, e.g., reaction times (Engelhardt et al., 2011), or eye-tracking (Grodner and Sedivy, 2011). We leave the question of under what conditions, if any, listeners might make informativity generalizations based on over-modification as an issue for future research. Nonetheless our results clearly demonstrate that prior beliefs of the listener about characteristics of referring expressions, and not just the optimality of an utterance with respect to a particular context, are an important factor for understanding reference generation and understanding. Along these lines, it will be important in future research to further investigate reasons why speakers might include more information in a referential expression than is strictly necessary for identifying a referent (for discussion see Isaacs and Clark, 1987; also Heller et al., 2012; Gorman et al., 2013; Gegg-Harrison and Tanenhaus, in review).

We began by considering how listeners might make rational use of linguistic information despite the fact that the linguistic input often includes more or less information than what is necessary and sufficient for a given referential intention. The current results help provide a critical piece of the puzzle: listeners can flexibly adapt their estimates of an expected amount of information associated with the given referential intention. Our findings demonstrate that the process of adaptation includes statistical inferences. Those inferences are conditioned on factors such as types of evidence (under- and over-informative), classes of adjectives, and listeners' prior beliefs about how reliably a particular type of non-optimal utterance would convey information about whether the talker would be non-optimal in the future.

Individual Differences

Although it was not a main focus of our study, the results in Experiments 2A and 2B revealed clear individual differences among participants with regard to their construal of the referential task. In Experiment 2B, we used an identical set of visual and audio stimuli as in Experiment 2A while drawing participants' attention to the fact that this is a task about evaluating the quality of instructions produced by two individual talkers. This manipulation made participants' responses significantly more uniform such that a larger proportion of participants provided responses that indicated informativity-based, rather than form-based, generalization. This suggests that participants vary in their construal of a task, a context, and a goal of referential communication even in a simple paradigm like the one we used in our study (for individual differences in semantic and pragmatic interpretations of utterances, see Noveck and Posada, 2003; Bott and Noveck, 2004; Degen and Tanenhaus, 2015;

Yildirim et al., 2016). Importantly, participants' assumptions about the task can determine the dimensions along which they generalize (see Brown-Schmidt and Fraundorf (2015) for evidence that perceived interaction influences use of common ground information).

Future Directions

The differences among participants suggest that one fruitful direction for future research will be to look at various contextual factors that likely influence the process of speaker-specific generalization of referential expectations. We mentioned in the introduction that studies on phonetic adaptation and generalization revealed that listeners structure their knowledge with respect to talker groups and situations. For instance, listeners do not *indiscriminately* generalize their knowledge about one talker's speech categories to a different talker, but facilitation after exposure to multiple talkers with the same foreign accent generalizes to new speakers with the same or similar accents (Bradlow and Bent, 2008). Similarly, listeners may be able to structure their expectations for referential expressions according to speaker groups or conversational contexts. For instance, adult speakers may produce more redundant modifiers when talking to a young child compared to when talking to another adult (e.g., *Look at the big brown doggy*! when there is only one dog in sight). Integrating contextual factors like this would help listeners "explain away" some of the variability observed within a speaker and further reduces the risk of under- or over-generalization.

We believe that our results have implications for research in reference production, including reference expression generation models (REG models). Models to date appear to take into account some manner of contextual information, primarily including referents that are visually or linguistically present in the context (see Krahmer and van Deemter, 2012, for a survey of work in REG to date). Some models attempt to accommodate interlocutor-specific information (e.g., Heeman and Hirst, 1995; Jordan and Walker, 2005) by producing referential expressions that reflect conceptual pact information (referential expressions that have been negotiated between particular interlocutors, see Clark and Wilkes-Gibbs, 1986; Brennan and Clark, 1996). We propose that future models of reference production should also take into account how interlocutors negotiate their referential expressions to find the most optimal level of reference given their certainty about the contextual and mutually shared information. In particular, such models should examine how these expectations might change over time given both the evidence at hand, and the interlocutors' prior beliefs.

Another fruitful line of research is examining how children treat under- and over-modifying utterances and whether they adapt their referential expectations in a talker-specific manner. Previous studies have reported that preschoolers can discriminate talkers' pragmatic abilities (e.g., Koenig and Harris, 2005; Scofield and Behrend, 2008), based on utterances with clear errors (e.g., using "key" to refer to a ball). It is, however, yet to be clear whether they can distinguish talkers based on the *quantity* of information provided (Eskritt et al., 2008). We have conducted a preliminary study using a paradigm similar to Experiment 1 in the current paper. We found that preschoolers, unlike adults, have difficulty associating under-modifying utterances with an individual talker (Pogue et al., 2015). This may be due to a number of possible reasons including their limited memory and attention span, general insensitivity to pragmatic principles in conversation and weaker assumptions for across-talker variability. Further investigation, both offline judgment studies like ours as well as online eye-movement studies, is necessary to paint a complete picture of the developmental trajectory of the ability to derive referential expectations.

Finally, our results open up a number of questions as to what is intended by *informativity*. As we discussed in the introduction, most theories so far have defined informativity as an expected amount of information with respect to an array of referents in a visual scene. Anything that exceeds the amount is considered over-informative and anything that falls short of it is considered under-informative. And these deviations are expected to trigger pragmatic inferences. Our results, however, yield strong support for the view that what counts as informative can change depending on a talker and a context. Listeners constantly update their expectations as they gain more information about the talker and the context. Future studies on informativity should therefore explore processes in which the speaker and the listener negotiate means and a context of reference, reducing uncertainty regarding form-referent mappings in a collaborative dialog.

AUTHOR CONTRIBUTIONS

AP, CK, and MT designed the research, AP performed the research, AP analyzed the data and drafted the paper. CK and MT provided feedback and edited AP's drafts.

ACKNOWLEDGMENTS

Thanks to T. Florian Jaeger, the HLP and Kurumada-Tanenhaus (KurTan) Labs for helpful feedback and advice, Matthew Overlan and Sarah Bibyk for voice recordings, and Chelsea Marsh and Olga Nikolayeva for help in creating stimuli and pilot testing.

FUNDING

This research was funded by NIH grant HD27206 awarded to MT and a NSERC PGS-D awarded to AP.

REFERENCES

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439. doi: 10.1006/jmla.1997.2558

Altmann, G., and Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition* 30, 191–238. doi: 10.1016/0010-0277(88)90020-0

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., and Trueswell, J. C. (2000). The rapid use of gender information: evidence of the time course of pronoun resolution from eyetracking. *Cognition* 76, B13–B26. doi: 10.1016/S0010-0277(00)00073-1

Arnold, J. E., Hudson Kam, C. L., and Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 914–930.

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011). Overspecification facilitates object identification. *J. Prag.* 43, 361–374. doi: 10.1016/j.pragma.2010.07.013

Aylett, M. P., and Turk, A. (2004). The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Lang. Speech* 47, 31–56. doi: 10.1177/00238309040470010201

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Belke, E. (2006). Visual determinants of preferred adjective order. *Vis. Cogn.* 14, 261–294. doi: 10.1080/13506280500260484

Belke, E., and Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: analyses of viewing patterns and processing times during "same"-"different" decisions. *Eur. J. Cogn. Psychol.* 14, 237–266. doi: 10.1080/09541440143000050

Benaglia, T., Chaveau, D., Hunter, D. R., and Young, D. S. (2009). mixtools: an R package for analyzing finite mixture models. *J. Stat. Softw.* 32, 1–29. doi: 10.18637/jss.v032.i06

Bott, L., and Noveck, I. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *J. Mem. Lang.* 51, 437–457. doi: 10.1016/j.jml.2004.05.006

Bradlow, A. R., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493.

Brown-Schmidt, S., and Fraundorf, S. H. (2015). Interpretation of informational questions modulated by joint knowledge and intonational contours. *J. Mem. Lang.* 84, 49–74. doi: 10.1016/j.jml.2015.05.002

Brown-Schmidt, S., and Hanna, J. (2011). Talking in another person's shoes: incremental perspective-taking in language processing. *Dialog Discourse* 2, 11–33. doi: 10.5087/dad.2011.102

Brown-Schmidt, S., and Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. *Cogn. Sci.* 32, 643–684. doi: 10.1080/03640210802066816

Brown-Schmidt, S., Yoon, S., and Ryskin, R. A. (2015). "People as contexts in conversation," in *Psychology of Learning and Motivation*, Vol. 62, ed. B. H. Ross (San Diego, CA: Elsevier Academic Press), 59–99.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Davies, C., and Katsos, N. (2009). "Are interlocutors as sensitive to over-informativeness as they are to under-informativeness?," in *Proceedings of the Workshop on Production of Referring Expressions: Bridging Computational and Psycholinguistic Approaches*, Amsterdam: PRE-Cogsci-09.

Davies, C., and Katsos, N. (2010). Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua* 120, 1956–1972. doi: 10.1016/j.cognition.2011.02.015

Degen, J., and Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cogn. Sci.* 39, 667–710. doi: 10.1111/cogs.12171

Deutsch, W., and Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition* 11, 159–184. doi: 10.1016/0010-0277(82)90024-5

Drager, K. (2010). Sociophonetic variation in speech perception. *Lang. Linguist. Compass* 4, 473–480. doi: 10.1111/j.1749-818X.2010.00210.x

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. (l995). Eye-movements as a window into spoken language comprehension in natural contexts. *J. Psycholinguist. Res.* 24, 409–436. doi: 10.1007/BF02143160

Engelhardt, P. E., Bailey, K. G. D., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity. *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Engelhardt, P. E., Demiral, S. B., and Ferreira, F. (2011). Over-specified referring expressions impair comprehension: an ERP study. *Brain Cogn.* 77, 304–314. doi: 10.1016/j.bandc.2011.07.004

Eskritt, M., Whalen, J., and Lee, K. (2008). Preschoolers can recognize violations of the Gricean maxims. *Br. J. Dev. Psychol.* 26, 435–443. doi: 10.1348/026151007X253260

Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336:998. doi: 10.1126/science.1218633

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 8051–8056. doi: 10.1073/pnas.1216438110

Gorman, K. S., Gegg-Harrison, W., Marsh, C. M., and Tanenhaus, M. K. (2013). What's learned together stays together: speakers' choice of referring expression reflects shared experience. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 843–853. doi: 10.1037/a0029467

Grice, H. P. (1975). "Logic and conversation," in *Syntax and Semantics: Speech Acts*, Vol. 3, eds P. Cole and J. L. Morgan (New York, NY: Seminar Press), 225–242.

Grodner, D., and Sedivy, J. (2011). "The effect of speaker-specific information on pragmatic inferences," in *Processing and Acquisition of Reference*, eds E. Gibson and N. J. Pearlmutter (Cambridge, MA: MIT Press), 239–272.

Hanna, J. E., Tanenhaus, M. K., and Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *J. Mem. Lang.* 49, 43–61. doi: 10.1016/S0749-596X(03)00022-6

Heeman, P. A., and Hirst, H. (1995). Collaborating on referring expressions. *Comput. Linguist.* 21, 351–382.

Heller, D., Gorman, K. S., and Tanenhaus, M. K. (2012). To name or to describe: shared knowledge affects choice of referential form. *Topics Cogn. Sci.* 4, 290–305. doi: 10.1111/j.1756-8765.2012.01182.x

Heller, D., Grodner, D., and Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition* 108, 831–836. doi: 10.1016/j.cognition.2008.04.008

Hertwig, R., and Gigerenzer, G. (1999). The "conjunction fallacy" revisited: how intelligent inferences look like reasoning errors. *J. Behav. Decis. Mak.* 12, 275–305. doi: 10.1002/(SICI)1099-0771(199912)12:4<275::AID-BDM323>3.3.CO;2-D

Isaacs, E. A., and Clark, H. H. (1987). References in conversation between experts and novices. *J. Exp. Psychol. Gen.* 116, 26–37. doi: 10.1037/0096-3445.116.1.26

Jaeger, T. F. (2010). Redundancy and reduction: speakers manage information density. *Cogn. Psychol.* 61, 23–62. doi: 10.1016/j.cogpsych.2010.02.002

Jordan, P. W., and Walker, M. (2005). Learning content selection rules for generating object descriptions in dialogue. *J. Artif. Intell. Res.* 24, 157–194.

Kleinschmidt, D. F., and Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar and adapt to the novel. *Psychol. Rev.* 122, 148–203. doi: 10.1037/a0038695

Koenig, M. A., and Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Dev.* 76, 1261–1277. doi: 10.1111/j.1467-8624.2005.00849.x

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Prag.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Krahmer, E., and van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI_a_00088

Kronmüller, E., and Barr, D. J. (2015). Referential precedents in spoken language comprehension: a review and meta-analysis. *J. Mem. Lang.* 56, 436–455.

Levy, R., and Jaeger, T. F. (2007). "Speakers optimize information density through syntactic reduction," in *Advances in Neural Information Processing Systems*

(NIPS), Vol. 19, eds B. Schlökopf, J. Platt, and T. Hoffman (Cambridge, MA: MIT Press), 849–856.

Metzing, C., and Brennan, S. E. (2003). When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions. *J. Mem. Lang.* 49, 201–213. doi: 10.1016/S0749-596X(03)00028-7

Nadig, A. S., and Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* 13, 329–336. doi: 10.1111/j.0956-7976.2002.00460.x

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *J. Lang. Soc. Psychol.* 18, 62–85. doi: 10.1177/0261927X99018001005

Noveck, I., and Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain Lang.* 85, 203–210. doi: 10.1016/S0093-934X(03)00053-1

Oaksford, M., and Chater, N. (2001). The probabilistic approach to human reasoning. *Trends Cogn. Sci.* 5, 349–357. doi: 10.1016/S1364-6613(00)01699-5

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 89–110. doi: 10.1515/ling.1989.27.1.89

Pogue, A., Kurumada, C., and Tanenhaus, M. K. (2015). Speaker-based generalization of quantity implicature in preschoolers. *Presentation Given at the 40th Annual Boston University Conference on Language Development*. Boston, MA.

Reinisch, E., and Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *J. Exp. Psychol. Hum. Percept. Perform.* 40, 539–555. doi: 10.1037/a0034409

Scofield, J., and Behrend, D. A. (2008). Learning words from reliable and unreliable speakers. *Cogn. Dev.* 23, 278–290. doi: 10.1016/j.cogdev.2008.01.003

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguist. Res.* 32, 3–23. doi: 10.1023/A:1021928914454

Sedivy, J. C. K., Tanenhaus, M., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition* 71, 109–147. doi: 10.1016/S0010-0277(99)00025-6

Sonnenschein, S. (1984). The effect of redundant communication on listeners: why different types have different effects. *J. Psycholinguist. Res.* 13, 147–166. doi: 10.1007/BF01067697

Strand, E. A., and Johnson, K. (1996). "Gradient and visual speaker normalization in the perception of fricatives," in *Proceedings of the Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference*, ed. D. Gibbon (Berlin: Mouton de Gruyter), 14–26.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychol. Rev.* 90, 293–315. doi: 10.1037/0033-295X.90.4.293

Viethen, J., Goudbeek, M., and Krahmer, E. (2012). "The impact of colour difference and colour codability on reference production," in *Proceedings of the 34th Annual Meeting of the Cognitive Science Society. Sapporo, Japan*, eds N. Miyake, D. Peebles, and R. Cooper (Austin, TX: Cognitive Science Society), 1084–1098.

Wolter, L., Gorman, K. S., and Tanenhaus, M. K. (2011). Scalar reference, contrast and discourse: separating effects of linguistic discourse from availability of the referent. *J. Mem. Lang.* 65, 299–317. doi: 10.1016/j.jml.2011.04.010

Xu, F., and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychol Rev.* 114, 245–272. doi: 10.1037/0033-295X.114.2.245

Yildirim, I., Degen, J., Tanenhaus, M. K., and Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *J. Mem. Lang.* 87, 128–143. doi: 10.1016/j.jml.2015.08.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 February 2016
doi: 10.3389/fpsyg.2016.00153

How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification

Paula Rubio-Fernández*

Centre for the Study of Mind in Nature, University of Oslo, Oslo, Norway

Color adjectives tend to be used redundantly in referential communication. I propose that redundant color adjectives (RCAs) are often intended to exploit a color contrast in the visual context and hence facilitate object identification, despite not being necessary to establish unique reference. Two language-production experiments investigated two types of factors that may affect the use of RCAs: factors related to the efficiency of color in the visual context and factors related to the semantic category of the noun. The results of Experiment 1 confirmed that people produce RCAs when color may facilitate object recognition; e.g., they do so more often in polychrome displays than in monochrome displays, and more often in English (pre-nominal position) than in Spanish (post-nominal position). RCAs are also used when color is a central property of the object category; e.g., people referred to the color of clothes more often than to the color of geometrical figures (Experiment 1), and they overspecified atypical colors more often than variable and stereotypical colors (Experiment 2). These results are relevant for pragmatic models of referential communication based on Gricean pragmatics and informativeness. An alternative analysis is proposed, which focuses on the efficiency and pertinence of color in a given referential situation.

Keywords: redundancy, color adjectives, object requests, informativeness, efficiency, pertinence, referential contrast

OPEN ACCESS

Edited by:
Emiel Krahmer,
Tilburg University, Netherlands

Reviewed by:
Ruud Koolen,
Tilburg University, Netherlands
Judith Degen,
Stanford University, USA

****Correspondence:***
Paula Rubio-Fernández
paula.rubio-fernandez@ifikk.uio.no

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *02 August 2015*
Accepted: *27 January 2016*
Published: *19 February 2016*

Citation:
Rubio-Fernández P (2016) How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. Front. Psychol. 7:153. doi: 10.3389/fpsyg.2016.00153

INTRODUCTION

Redundancy is generally defined in terms of informativeness: to say that an expression is redundant is to say that it is over-informative or overspecific (Engelhardt et al., 2006; Sedivy, 2007; Davies and Katsos, 2010; Arts et al., 2011a,b). According to this view the following utterances are redundant:

(a) ? John is a bachelor and he is unmarried.
(b) ? Today we are meeting at 7 pm in the evening.
(c) ? Give me the blue cup (uttered in a situation where there is only one cup).

While the first two examples are redundant because they are repetitive (e.g., a bachelor is unmarried by definition), the last example is redundant because it includes a non-contrastive use of a color adjective (i.e., 'blue' is not used to distinguish the intended cup from another cup of a different color). This paper focuses on the last type of redundant expressions; namely, redundant color adjectives (RCAs) in object requests. Unlike other types of speech acts involving reference, object requests require that the hearer visually identify the object in the physical environment as part of the pragmatic process of reference assignment. This feature of object requests makes them ideal for a pragmatic investigation of the role of visual processes in the production of referential expressions – which is the aim of the present study.

According to Grice's (1975) Maxim of Quantity, speakers should try to provide their interlocutors with as much information as they need, but not more. Thus, in a situation where there is only one cup, the unmodified referential expression 'the cup' should be preferred to 'the blue cup,' other things being equal. Contrary to this theoretical expectation, experimental research has shown again and again that people tend to use adjectives redundantly in referential communication (e.g., Pechmann, 1989; Sedivy, 2003; Maes et al., 2004; Koolen et al., 2013). Another recurrent finding in the literature is that color adjectives tend to be used redundantly more often than other types of adjectives, especially relative adjectives such as 'large' or 'small' (Pechmann, 1989; Belke and Meyer, 2002; Nadig and Sedivy, 2002; Arts et al., 2011a,b).

The study reported in this paper investigated what factors affect the production of RCAs as a way to understand why they are so frequently used in object requests. Before I turn to that issue, I will address a theoretical question that has often been discussed in the pragmatics literature on referential communication: is color encoded because it is salient for the speaker or for the hearer? This question is important for the pragmatic analysis of color overspecification that I will propose next, which is based on efficiency.

The Two Sides of Color Salience

As with other pragmatic aspects of reference production (e.g., articulatory attenuation; Brennan et al., 2010), it has often been discussed whether overspecification is a 'speaker-internal' or 'hearer-oriented' process (Arnold, 2008). Some authors have suggested that using adjectives redundantly may be easier for the speaker because it precludes the need to determine whether or not a certain adjective is necessary for unique reference (Pechmann, 1989; Belke and Meyer, 2002; Belke, 2006; Engelhardt et al., 2006; Koolen et al., 2013). It has also been argued that an overspecified description may help the hearer identify the intended object (Sonnenschein and Whitehurst, 1982; Mangold and Pobel, 1988; Nadig and Sedivy, 2002; Paraboni et al., 2007; van der Sluis and Krahmer, 2007; Arts et al., 2011a,b).

In the case of attenuation, it has been argued that attenuating the articulation of a word that is predictable in the context (vs. a word encoding new information) may be easier for the speaker insofar as it requires less articulatory effort than pronouncing it clearly (as is often done with new information). It is therefore possible that articulatory attenuation is simply easier for the speaker and benefits the hearer's comprehension only fortuitously (but cf. Galati and Brennan, 2010). The case of overspecification is somewhat different, however, since identifying a property of a referent and encoding it in an utterance is generally harder (or more costly) for the speaker than not doing so. Since overspecification happens precisely in contexts where the encoded property is not necessary to establish unique reference, speakers' choice of a longer referential expression needs to be explained.

One way in which overspecification may be easier for the speaker is by eliminating the search for potential competitors (i.e., objects of the same category as the intended referent) in the visual display. Pechmann (1989) observed that speakers often started producing overspecific referential expressions before they had finished scanning a display, suggesting redundancy may indeed facilitate reference production for the speaker. It must be noted, however, that this kind of evidence only explains the use of redundant adjectives in relatively large displays where scanning would be time consuming, but not in sparse displays where the speaker could determine at a glance that all objects are different.

However, even when overspecification would save the speaker the time to scan a display for potential competitors, such behavior would not be only 'for the speaker', but also 'for the hearer.' Thus, if a speaker's referential strategy is to use a modified expression to preempt a possible ambiguity in a large display, then that default strategy is in itself evidence of audience design (while also being economical for the speaker). By contrast, a truly 'egocentric' speaker who was insensitive to the hearer's perspective would not bother scanning the display or specifying a property of the intended referent in case there was a competitor: a self-centered speaker who made an object request would simply produce a bare definite description and leave it up to the hearer to ask for more information or make a guess (in the event that the request turned out to be underspecific).

I want to argue further that, at least in face-to-face communication, trying to decide whether overspecification is for the speaker or for the hearer is pretty much futile. Given that in face-to-face communication the physical environment is part of the common ground between the speaker and the hearer (Clark and Marshall, 1981), what is salient for the speaker (e.g., the color of a cup) will generally be salient for the hearer as well. Most importantly, a Gricean speaker is entitled to assume that much when producing a referential expression. In other words, since speakers and hearers rely on the same perceptual mechanisms, a cooperative speaker is entitled to assume that anything that is perceptually salient to him will also be salient for his interlocutor when they share a physical environment.

Experimental pragmatics studies have repeatedly found that the interlocutors' sharing of a physical environment (what is known as 'co-presence') affects referential communication. For example, speakers' eye gaze can be used by hearers to assign reference to a linguistic expression in face-to-face communication (e.g., Richardson and Dale, 2005; Hanna and Brennan, 2007; Neider et al., 2010). Likewise, bearing in mind the goal of the task at hand can also help hearers disambiguate referring expressions in interactive games (e.g., Chambers et al., 2002, 2004; Hanna and Tanenhaus, 2004). Co-presence can also affect language production, as when speakers tell stories to interlocutors who either share a picture of the story with the speaker, or rely entirely on the speaker's narrative. In the latter condition, speakers tend to specify atypical objects more often than when these objects are visible to both interlocutors (Lockridge and Brennan, 2002). In this study I will argue that co-presence is relevant for the use of RCAs in object requests insofar as RCAs may facilitate object identification for the hearer.

In the remainder of this paper I will not try to discern whether speakers use RCAs in object requests because (a) they themselves find the color of the referent salient, or because (b) their interlocutor must find the color of the referent salient. Since interlocutors in face-to-face communication can normally assume that if (a) then (b), the speaker's and hearer's perspectives do not differ enough to be disentangled experimentally in such situations (for discussion, see Keysar, 1997; Brennan et al., 2010). Instead, I will treat reference as a 'collaborative process' between interlocutors (Clark and Wilkes-Gibbs, 1986) and try to argue that overspecification may be *efficient* in face-to-face communication.

In formulating an object request, the speaker's goal is to get the hearer to identify the object in the physical environment and, assuming the hearer is willing to comply with the request, both interlocutors come to share the same goal. Object requests in face-to-face communication are therefore an ideal test case for the view that referential communication requires verbal and visual coordination between interlocutors, from which it follows that some referential expressions may be more efficient than others.

Informativeness vs. Efficiency

Unlike computational psycholinguistics studies of reference production (e.g., Paraboni et al., 2007; Arts et al., 2011b; Koolen et al., 2013; Westerbeek et al., 2015), pragmatic accounts of referential communication have thus far failed to take into account perceptual factors in referential communication. For example, Sedivy (2003, 2004, 2007; Grodner and Sedivy, 2011) proposed a pragmatic analysis of color adjectives based on 'default descriptions,' according to which the default description of variable-color objects (e.g., a cup) includes a color adjective, while the default description of stereotypical-color objects (e.g., a banana) does not include a color adjective. This distinction explains why requests for variable-color objects tend to include RCAs, while requests for stereotypical-color objects only include color adjectives if there is a competitor in the display (e.g., a green banana; Sedivy, 2003, 2004; see also Westerbeek et al., 2015).

However, Sedivy's account does not take perceptual factors into account, even though the physical environment is part of the common ground between interlocutors in face-to-face communication. According to Sedivy's model, the referential expression 'the blue cup,' for example, would be redundant or 'non-contrastive,' if there is only one cup in the display. However, if we consider visual object identification as part of the pragmatic process of reference resolution in object requests, a pragmatic analysis of the expression 'the blue cup' must take into account not only the number of cups in the display, but also the colors of the other objects. Compare in this respect a visual search for a blue cup in a display where the cup is the only blue object, with the same visual search when all the objects in the display are blue. According to the standard pragmatic view, the referential expression 'the blue cup' would be equally over-informative in both contexts (so long as there was only one cup in each display). However, in the analysis I am proposing, the same referential expression would not be equally *inefficient*, since knowing the color of the cup would facilitate object identification in the polychrome display but not in the monochrome display.

Contrary to previous accounts, I want to propose that a pragmatic analysis of referential communication needs to be cast in terms of *efficiency* rather than *informativeness*. As was explained in the introduction, redundancy is traditionally described in terms of informativeness. However, such an analysis is only appropriate for statements, whose goal is to inform the hearer of a state of affairs (e.g., 'It's raining'); object requests, by contrast, are not informative as such. In terms of efficiency, a linguistic expression would be redundant if there was a more succinct alternative that would have achieved the goal of the speech act equally well. Given the goal of an object request, an optimal referential expression in an object request is one that allows the hearer to identify the intended object in the most efficient way. According to this view, RCAs should be understood as more or less efficient in a given context rather than being necessarily considered pragmatically infelicitous for being over-informative (Engelhardt et al., 2006, 2011).

An account of referential communication in terms of efficiency has the advantage that efficiency is a finer-grained notion than the standard three-way distinction between 'underspecific,' 'minimal' and 'overspecific referential expressions,' which has characterized Gricean analyses so far (e.g., Heller et al., 2012; Pogue et al., 2015). First of all, an efficiency-based analysis must take into account the specificity of a referential expression, since an underspecific referential expression (e.g., asking for 'the cup' in a situation where there are two cups) is less efficient than a minimal referential expression that establishes unique reference (e.g., 'the blue cup' in the same situation) insofar as the former expression leaves the hearer to choose randomly or ask for clarification.

In addition, looking at efficiency allows a deeper analysis of referential overspecification. For example, referring to the only cup in a display as 'the blue cup' would be more efficient if the cup was the only blue object than if there was also a blue jug in the display. However, color distinctiveness is not the only factor that may affect the relative efficiency of a referential expression: the number of objects in the display is also relevant. Thus, mentioning the color of the cup in a display of four objects would not be very efficient if two of them were blue, but the same expression would be considerably more efficient if the two blue objects were among 10 other objects of a different color. An analysis of referential communication based on efficiency is therefore much finer-grained than standard analyses in terms of informativeness. In this respect, while the idea that color may facilitate object identification is hardly new (e.g., Sonnenschein and Whitehurst, 1982; Mangold and Pobel, 1988; Paraboni et al., 2007; Arts et al., 2011a), the proposal to analyse RCAs as more or less efficient in a given context is novel and departs, in important ways, from standard pragmatic analyses in terms of informativeness.

The Two Sides of Efficiency

An efficient referential expression is one that facilitates the identification of the intended referent for the hearer relative to other referential expressions. In the case of RCAs in object requests, a direct measure of efficiency would require comparing the speed of hearers' identification of the referent following minimal and modified instructions (e.g., 'Give me the cup' vs. 'Give me the blue cup' in the same display containing only one cup). In a recent eye-tracking study investigating this particular question, I collected continuous eye-tracking measures of target identification and response times, and found an advantage for the modified instructions (containing RCAs) in all measures and across all conditions (which included different types of visual display; see Rubio-Fernández, under review). The results of this study therefore confirm that redundancy can be efficient, contrary to what standard pragmatic models have assumed to date.

Hearers' eye movements and response times provide a direct measure of efficiency in referential communication when comparing modified and unmodified instructions. However, comparative comprehension data are not available to speakers and therefore do not inform their choice of referential expression. In that sense, comprehension data provide only half the picture – as one would expect. When speakers formulate an object request in face-to-face communication, what they have at their disposal is visual information about the environment in which their interlocutor must identify the target object (including the object's contrastive properties). Therefore, reference production studies must also be carried out in order to establish whether speakers use RCAs when they can gauge that it may be efficient for their interlocutor in the visual context.

For example, a speaker who produces an efficient object request should be sensitive to the density of the display from which their interlocutor must select the referent. More specifically, such a speaker should have a stronger tendency to produce RCAs the denser the display is with objects, since that increases the difficulty of the hearer's visual search. The tendency to provide our interlocutors with more information when they are looking for an object in a cluttered environment is generally an efficient strategy, which can be investigated in a language-production study without having to measure the speed of the interlocutor's response (see Paraboni et al., 2007; Clarke et al., 2013; Rubio-Fernández, under review).

In line with the above arguments, the present study only investigated factors affecting the production of RCAs in relation to the potential efficiency of such uses in the given visual context. I was therefore not concerned with the actual effect that using (or not using) RCAs may have on reference resolution (since such differential effects do not inform a speaker's choice of referential expression in the first place). In this sense, I will only consider efficiency from the viewpoint of the speaker: an efficient referential expression is one that the speaker *could reasonably expect* to help the hearer identify the intended referent in the visual context. This pragmatic notion of efficiency is broadly related to the speaker's cooperative intention, and is not dependent on whether the referential expression is actually effective for the hearer.

Factors Affecting the Use of Redundant Color Adjectives in Object Requests

It has been suggested before that factors other than considerations of unique reference may affect the choice of an adjective in an object request; for example, high-frequency adjectives and adjectives for salient properties are likely to be used in definite descriptions (Pechmann, 1989; Sedivy, 2004; Koolen et al., 2011). The present study investigated two types of factors that may affect the production of RCAs: visual-contextual factors and semantic-category factors.

Visual-contextual factors affect the use of RCAs in relation to the efficiency of color in a given situation; that is, to the extent that color may help the interlocutor identify the intended object. Two specific hypotheses were tested in relation to visual-contextual factors: first, RCAs are more efficient in an object request if the objects in the display are of different colors than if they are all the same color, especially if the referent is the only object of its color. This is so because color can be used to identify the intended referent in a polychrome display, but not in a monochrome one. I therefore predicted that more RCAs would be produced in polychrome displays than in monochrome displays. Such a difference was reported by Belke and Meyer (2002) and Koolen et al. (2013), although not in connection with the hearer's visual search for the referent.

Second, since language processing is incremental, color adjectives are a more efficient cue to the hearer's visual search in pre-nominal position than in post-nominal position. Eye-tracking studies have shown that a spoken instruction guides the hearer's eye movements incrementally (Spivey et al., 2001; Reali et al., 2006; Clarke et al., 2015). Thus, when an English interlocutor processes the overspecific instruction 'Give me the *blue* cup,' she uses the color adjective to guide her visual search for the cup. In contrast, when a Spanish interlocutor processes the overspecific instruction 'Dame la taza *azul,*' she starts looking for the cup before she gets to process the color adjective, possibly finding the referent without using its color as a cue. Therefore, even if adjective position is a syntactic constraint, it affects the hearer's visual search for the referent – hence its classification as a visual-contextual factor in this study. Given the difference in efficiency between pre-nominal and post-nominal RCAs, I predicted that more RCAs would be produced in English than in Spanish.

Semantic-category factors affect the use of color adjectives in relation to the noun that they modify, and hence according to our world knowledge of the category. For example, Sedivy (2003, 2004) found that color adjectives are used redundantly in requests for objects of variable colors (e.g., cups) but not of stereotypical colors (e.g., bananas). Two hypotheses were tested in relation to semantic-category factors: first, following up on Sedivy's findings, the present study investigated the use of RCAs for objects of atypical colors (e.g., a pink banana). It was predicted that RCAs would be used more often for atypical- than for variable- and stereotypical-color objects, since atypical-color objects violate our expectations about a given category. This hypothesis was recently supported by the results of Westerbeek et al. (2015), who modified the color of fruits and vegetables (which normally have stereotypical colors) in a referential communication study.

Second, I propose that color is more important or pertinent for some semantic categories than for others (e.g., clothes and cars, on the one hand, vs. geometrical figures and tools, on the other) and predict that people will produce more RCAs when color is pertinent for a given semantic category. The pertinence of color for a given category should have an effect on the frequency with which color adjectives are used to refer to that category, as suggested by collocations with nouns for which color is a central property (e.g., 'little black dress,' 'black tie,' 'white collar workers' or 'red sports car'). Underlying such frequency effects, however, it is possible that both speakers and hearers recognize an optimal level of description for any given category.

Just as referring to one's pet as 'my dog' is normally more appropriate than as 'my animal' (Reiter, 1991; Geurts, 2010), it is not unlikely that specifying the color of a certain object is generally appropriate on the grounds that color is a central property for the category (e.g., 'the black pen' vs. 'the black radio'). In such instances of color overspecification, Dale and Reiter (1995) have suggested that speakers may use 'reference scripts' that determine which properties are expected for a certain semantic category (for the related notion of 'default descriptions,' see Sedivy, 2003, 2004; Grodner and Sedivy, 2011). Thus, even if color might not necessarily be efficient in the visual context, it could be argued that specifying the color of clothes and shoes, for example, is generally pertinent for the requested object and therefore acceptable (and maybe even expected by the interlocutor, according to their reference script).

EXPERIMENT 1

The first experiment in the study investigated two hypotheses related to the efficiency of RCAs in object requests. First, whether speakers would produce more RCAs in polychrome displays in which the referent was the only object of its color, than in monochrome displays where all objects were the same color as the target. This pattern of results was previously observed by Belke and Meyer (2002) and Koolen et al. (2013). However, these studies investigated other factors in addition to the number of colors in the display (e.g., the effect of size and orientation contrast in the visual display). It is therefore not possible to establish to what extent these results were due to the effect of color contrast. In fact, Koolen et al. (2013) argue that it is 'scene variation' (generally understood as the number of dimensions along which the objects of a display may vary) which drives the use of RCAs in their study, and not color contrast *per se*.

There is also a methodological reason why the results of Belke and Meyer (2002) and Koolen et al. (2013) may not be conclusive: in both studies monochrome and polychrome trials were interspersed and it is therefore possible that the RCAs that were observed in monochrome trials were a carry-over effect from previous polychrome trials. Belke and Meyer (2002), for example, observed that color adjectives were overspecified in monochrome displays of geometrical figures up to 66.5% of the time. However, it is an open question whether their participants would have produced such a high proportion of RCA had they been exclusively presented with monochrome displays. The results of Rubio-Fernández (under review) suggest otherwise, since participants produced zero rates of RCAs when they requested geometrical figures from monochrome displays alone (for a recent investigation of consistency in referential overspecification, see Tarenskeen et al., 2015). Experiment 1 is therefore the first study to specifically investigate the effect of color contrast on the use of RCAs.

The second hypothesis to be investigated in relation to the efficiency of RCAs was whether English speakers would produce more RCAs than Spanish speakers, despite both languages having the same basic color terms. I have recently argued that, in face-to-face referential communication, color adjectives guide an interlocutor's visual search for the referent (Rubio-Fernández, under review). In this view, color adjectives are a more efficient cue in pre-nominal position than in post-nominal position because in the latter case the hearer's visual search is initially guided by the noun (and not by the color adjective). It was therefore hypothesized that RCAs would be produced more frequently in English than in Spanish.

The relative efficiency of color adjectives with regards to the incrementality of language processing is best investigated in relatively sparse displays, such as the ones used in Experiment 1. Thus, in a 4-object display, a Spanish hearer may be able to identify the intended referent in processing the noun, thus rendering the post-nominal color adjective useless as a visual cue. The incrementality of language processing is therefore an important factor in the production of RCAs across languages. However, this factor has not been previously investigated either in computational or in pragmatic studies on referential communication.

In order to test the above hypotheses, I designed the Paper Dolls task: a simple referential communication task in which participants had to ask the experimenter to click on paper clothes and shoes in a series of 4-garment displays following a model paper doll. This task also served to test a third hypothesis related to the effect of color pertinence on the production of RCAs: I predicted that both English and Spanish participants in the Paper Dolls task would produce a relatively high proportion of RCAs because color is a central property of clothes and shoes in Western cultures and may therefore feature in reference scripts for such categories. Consider in this respect how color coordination is important when we choose clothes and how some colors are even more fashionable than others, depending on the season. These effects, however, are not observable in all man-made objects, despite the fact that artifacts often come in different colors (e.g., Kitchenware or office supplies). One would therefore expect that the association between color and clothes should be stronger than the association between color and other types of artifacts.

Method

Participants

Thirty-nine undergraduate students from University College London and the University of Kent (UK), all native speakers of English (20 female), and 39 undergraduate students from the Universities of Oviedo and Baleares (Spain), all native speakers of Spanish (25 female), took part in the experiment for monetary compensation. All participants gave consent to have their voice recorded during the experiment. Ethics approval was obtained from University College London and the University of Baleares. Permission to run the experiment was obtained from all departments where data was collected.

Materials and Procedure

Six images showing a paper doll were designed so that each doll wore three garments of different colors (see **Figure 1**). Three displays of four paper clothes were constructed for each paper doll, with only one garment corresponding with what the paper doll was wearing (see **Figure 2**). In the polychrome condition, the displays included garments of four different colors. The same displays were used in the monochrome condition, only that all the garments were the same color as the target. Since the model paper dolls always wore 3 garments of different colors, color changed across trials in the monochrome condition (e.g., for the model doll in **Figure 1**, the monochrome displays were pink, blue, and brown). Target garments were the following colors: blue, yellow, green, red, pink, purple, orange, and brown.

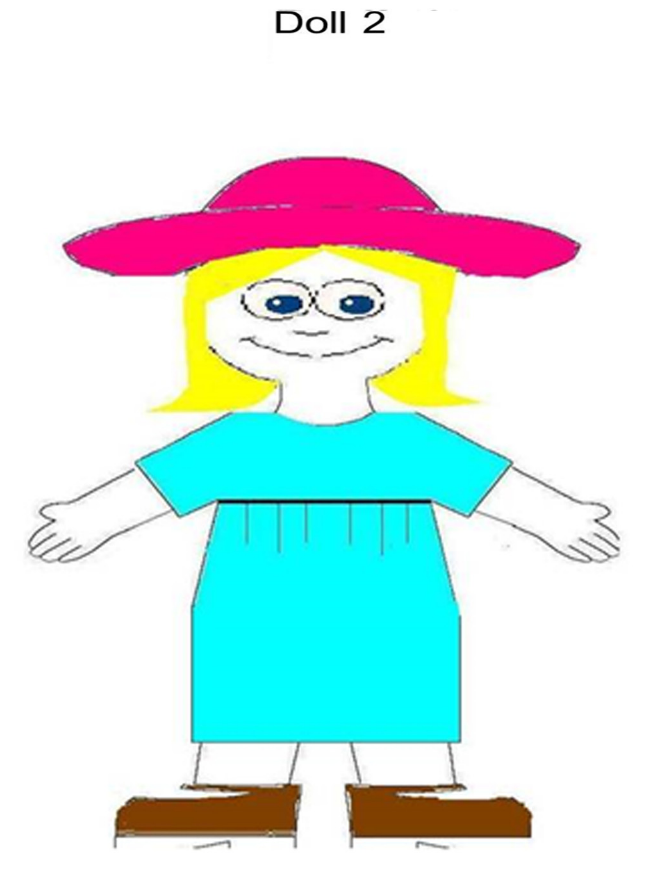


FIGURE 1 | Model paper doll used in both the monochrome and polychrome conditions.

Display type (Polychrome vs. Monochrome) and Language (English vs. Spanish) were manipulated between participants in a 2 × 2 design. The paper dolls were printed in color on A4 paper while the 4-garment displays were presented on a computer monitor using E-Prime. Given the simplicity of the task, participants were told that they were a control group in a study investigating the development of children's communicative skills. The aim of the original study was to see how pre-school children performed in an interactive game in which they had to dress a paper doll following a model and asked the experimenter for the paper clothes they needed. Adults were going to be administered the same task as the children in order to obtain control data to evaluate children's performance. The only difference with the children's task was that, instead of playing with cut-out dolls and real paper clothes, adults would be shown paper clothes on a computer monitor and the experimenter would click on their garment of choice in each display.

The experimenter waited until each instruction was completed to click on the designated target (as performing faster may invite Spanish speakers not to encode post-nominal color adjectives). Participants' requests were recorded and later coded by the experimenter as including or not including a RCA. Only referential expressions including both an adjective and a noun (e.g., 'The blue dress' or 'El vestido azul') were coded as overspecific.

Results

The data from both Experiments 1 and 2 were analyzed using non-parametric statistics because they were not normally distributed (which made parametric tests such as ANOVA and *t*-test unsuitable) and because the extreme data values observed in some conditions interfered with model convergence when mixed-model analyses were attempted.

Participants instructions conformed to the minimal or color-overspecific descriptions that were elicited (e.g., 'The dress' or 'The blue dress'). The mean proportions of RCAs for each Language and Display condition in the Paper Dolls task are shown in **Table 1**. A Kruskal–Wallis test was conducted on the proportions of RCAs, revealing a significant difference among conditions, $H(3) = 41.9, p < 0.001$.

Looking first at the effect of Display type, Mann–Whitney tests were carried out on participants' RCA scores in each language, revealing a significant effect of Display in English, $U = 71.5$, $Z = 3.32$, $p < 0.001$; and in Spanish, $U = 12.5$, $Z = 4.97$, $p < 0.001$, with RCAs being produced more often in polychrome displays than in monochrome displays in the two languages. In addition, of the 40 participants who took part in the polychrome version of the task, 19 participants (17 English speakers) used RCAs systematically, whereas of the 38 participants in the monochrome version, only six participants (all English speakers) used RCAs systematically. A Chi-square test corrected for continuity revealed that the difference between the number of participants who used RCAs systematically (and not systematically) in each type of display is significant, $X^2(1,N = 78) = 7.60$, $p < 0.006$, with more systematic uses of

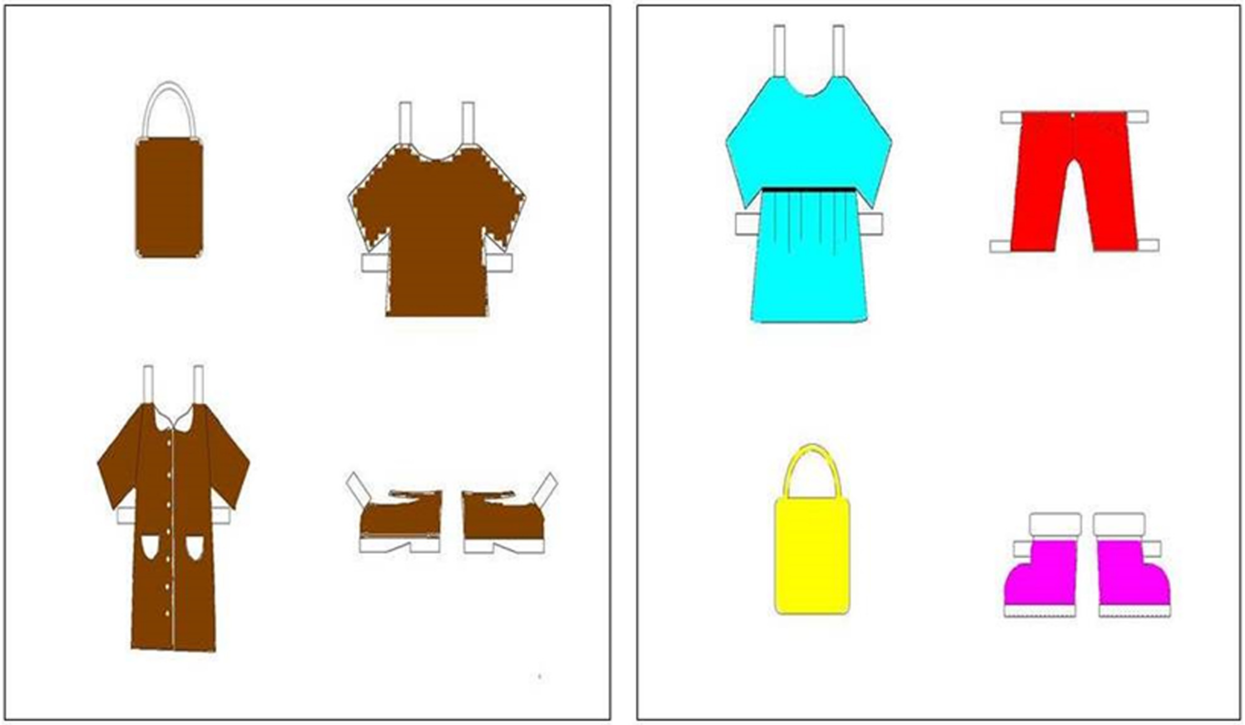


FIGURE 2 | Displays of paper clothes and shoes from the monochrome condition (left: four different garments of the same color) and the polychrome condition (right: four different garments of four different colors).

RCAs being observed in the polychrome condition than in the monochrome condition.

Looking at the effect of Language, Mann–Whitney tests were carried out on participants' RCA scores in each type of display, revealing a significant effect of Language in the Polychrome condition, $U = 42$, $Z = 4.26$, $p < 0.001$; and a marginally significant effect in the Monochrome condition, $U = 114$, $Z = 1.93$, $p = 0.054$, with RCAs being produced more often in English than in Spanish in both types of display. In addition, of the 39 English speakers who took part in the study, 23 used RCAs systematically, while of the 39 Spanish speakers only two used RCAs systematically. A Chi-square test with continuity correction revealed that the difference between the number of participants who used RCAs systematically (and not systematically) in each language is significant, $X^2(1,N = 78) = 23.6$, $p < 0.001$, with English speakers systematically producing RCAs more often than Spanish speakers.

Discussion

The results of Experiment 1 confirmed that visual-contextual factors affect the production of RCAs: first, RCAs were produced more often in polychrome displays than in monochrome displays, confirming that speakers tend to choose efficient referential expressions when formulating object requests. This pattern of results replicates previous findings by Belke and Meyer (2002) and Koolen et al. (2013), who observed a higher proportion of RCAs in polychrome displays than in monochrome displays. However, a direct comparison with the proportions of RCAs observed in those studies would not be reliable, since they used different types of objects and manipulated a number of other factors (e.g., size contrast and orientation), which may have also affected their results together with the effect of color contrast.

Second, RCAs were produced more often in English than in Spanish, suggesting again that speakers are efficient in their use of RCAs since pre-nominal color adjectives are a more efficient cue to the hearer's visual search than post-nominal color adjectives (for relevant visual-search studies, see Spivey et al., 2001; Reali et al., 2006; Clarke et al., 2015).

Regarding semantic-category factors, participants in the Paper Dolls task revealed a strong tendency to use RCAs when referring to clothes and shoes, both in the English- and Spanish-Polychrome conditions. In order to evaluate the magnitude of this effect, I will compare the results of Experiment 1 with those reported in Rubio-Fernández (2015). In the latter study, which I conducted in parallel with the present one, I used the Figures and Stickers task: a similar test to the Paper Dolls task in which participants had to ask the experimenter to click on a geometrical figure in a series of 4-figure displays following a model figure. A comparison between these two studies is reliable for two reasons: first, the materials and procedures of the two tasks were identical, with the exception of the shapes used in the displays. Second, color is a central property of clothes and shoes, whereas it is not a particularly central property of geometrical figures. It

TABLE 1 | Mean proportions of redundant color adjectives (SD) produced in the Paper Dolls task.

Display	Paper Dolls	
	English	Spanish
Polychrome	0.95 (0.15)	0.59 (0.36)
Monochrome	0.37 (0.49)	0.05 (0.17)

therefore follows from my prediction regarding color pertinence that English and Spanish speakers should produce more RCAs when performing the Paper Dolls task than when performing the Figures and Stickers task.

The results of Rubio-Fernández (2015) revealed that English speakers produced RCAs 46% of the time in polychrome displays of geometrical figures (*SD*=5.06), while Spanish speakers did so 14% of the time (*SD*=3.03). Relative to the data elicited with geometrical figures, English speakers produced more than twice as many RCAs when referring to clothes and shoes in Experiment 1 (0.46 vs. 0.95), while Spanish speakers did so four times as often (0.14 vs. 0.59). The comparison between the results of the Paper Dolls task and the Figures and Stickers task used by Rubio-Fernández (2015) confirm that people tend to produce RCAs when color is a central property of the noun category, as it is the case with clothes and shoes.

Rubio-Fernández (under review) also tested English participants on the Figures and Stickers task using monochrome displays of geometrical figures. The comparison between the monochrome conditions of the Figures and Stickers task and the Paper Dolls task is critical for the present investigation since the two factors considered in this study (i.e., visual-contextual factors and semantic-category factors) are at odds in those conditions. The question is therefore whether English speakers would produce more RCAs when referring to clothes than to geometrical figures in monochrome displays. If that is the case, semantic-category factors would trump visual-contextual factors since specifying the color of a pair of shoes in a monochrome display, for example, would not facilitate the identification of the shoes for the hearer.

Rubio-Fernández (under review) reported that English speakers produced zero rates of RCAs when referring to geometrical figures in monochrome displays. Relative to these results, the proportion of RCAs observed in the English-Monochrome condition of the Paper Dolls task (0.37) was significantly higher (Mann–Whitney test, $U = 95$, $Z = 2.48$, $p < 0.014$). This pattern of results is revealing, since color does not facilitate object identification when all objects are the same color and therefore, the use of color adjectives to refer to clothes in the monochrome condition was driven by the semantic category of the noun. This effect, however, was only observed in English, with the rates of RCAs produced in the Spanish-Monochrome condition of the Paper Dolls task being close to zero. This is also an interesting difference, since there is no reason to suppose that color is more pertinent for clothes in British than in Spanish culture.

The picture emerging from Experiment 1 is therefore a complex one, with color contrast (monochrome vs. polychrome), adjective position (pre-nominal vs. post-nominal) and semantic category of the noun (clothes and shoes vs. geometrical figures) having a combined effect on the production of RCAs. The interaction of these factors suggests that the production of RCAs is highly context-dependent and requires a finer-grained analysis than a standard evaluation of informativeness. According to such pragmatic analyses, the RCAs observed in the various conditions discussed above would all have been equally over-informative, yet the variability in the data would remain unaccounted for unless other factors were taken into consideration.

EXPERIMENT 2

In addition to the effect of color pertinence investigated in Experiment 1, Experiment 2 investigated a second semantic-category factor; namely the effect of color typicality. More specifically, whether people would use RCAs to refer to objects of atypical colors. Sedivy (2003, 2004) observed that people tend to use RCAs for objects of variable colors (e.g., a blue cup) but not for objects of stereotypical colors (e.g., a yellow banana). Regarding objects atypical colors (e.g., a pink banana), Westerbeek et al. (2015) have recently shown that color tends to be overspecified more often when it is atypical of an object than when it is typical (for a study of shape and material typicality, see also Mitchell et al., 2013). In addition, and relevant to the present argument that RCAs can be efficient in a given visual context, Westerbeek et al. (2015) found that the preference for atypical colors was stronger when color was more important for object identification.

Westerbeek et al. (2015) mainly used displays of fruits and vegetables (although their second experiment also included other stereotypical-color objects) and presented them in more or less typical colors. More in line with the various types of objects used by Sedivy (2003, 2004), Experiment 2 used objects of stereotypical, variable and atypical colors (e.g., an orange carrot, a red bicycle and a pink banana). The aim of Experiment 2 was to investigate the overspecification of atypical colors as a test of the view that RCAs can be efficient, and therefore cooperative in nature (i.e., a test of the pragmatics of color overspecification). For this purpose and unlike the above studies, I manipulated not only color typicality, but also the type of instruction that the participants received at the start of the task (i.e., standard vs. cautionary instructions, with the latter alerting participants to the possibility of a communication breakdown).

There are at least two possible reasons why a speaker may choose to overspecify an atypical color. The first reason would be a bottom-up effect resulting from a violation of the speaker's word knowledge. For example, a pink banana would be such an odd banana that its color might be highly salient and therefore mentioned in a request for the object even if unnecessary for unique reference. However, there is also a compatible, top-down process by which a speaker would mention the atypical color of an object in order to prevent the hearer from deriving the wrong presupposition. For example, if the speaker wanted a pink banana from among various objects but did not mention its color, the hearer would probably start looking for a yellow object. This hypothesis is supported by the results of a visual-world study by Huettig and Altmann (2011), which showed that when people hear the name of a stereotypical-color entity (e.g., 'spinach', which is typically green), they fixate on objects of that color, even though the actual color of the category was not mentioned. Thus, in order to spare the hearer unnecessary effort, the speaker might choose to use a RCA when referring to an atypical-color object.

This second factor would be a pragmatic factor and is related to the question of whether speakers may be cooperative when they use RCAs in their object requests. It is important to note that these two factors are compatible and, in fact, the second, top-down factor depends on the first, bottom-up factor: in order for the speaker to want to spare the hearer unnecessary effort, he must have first detected that the color of the target object violated his world knowledge of the category. Therefore, the aim of Experiment 2 was not to investigate which of these two factors plays a role in the production of RCAs. Instead, the aim of the experiment was to investigate whether speakers may go beyond noticing that the color of a certain object is atypical, and encode RCAs to facilitate object identification for the hearer.

In order to investigate this question, I designed the Yellow Pig task, a simple referential communication task in which participants had to ask the experimenter to click on a target in a series of 4 × 4 displays. In order to investigate the effect of color typicality, the targets were stereotypical-color fruits, vegetables, and animals (e.g., a brown dog); atypical-color fruits, vegetables, and animals (e.g., a pink banana) and variable-color artifacts (e.g., a silver toaster). The latter condition served as a neutral baseline for color typicality, understood as the midpoint between the stereotypical and atypical conditions.

Regarding the question of whether participants are being cooperative when they mention atypical colors, a manipulation was introduced in the instructions intended to make participants more sensitive to a potential communication breakdown between the participant and the experimenter. Participants in the Cautionary condition were made to believe that participants in the pilot phase of the study had sometimes failed to specify which figure was the target and the experimenter had had to ask which of two possible referents they were referring to. Importantly, ambiguity was never an issue in the actual test (with the displays always including different types of figures).

The key hypothesis was that, if participants mentioned atypical colors in order to spare the experimenter unnecessary effort in her object search, a subtle manipulation in the instructions should result in an increase in RCAs in the atypical-color condition but not necessarily in the other two conditions, since only the atypical-color condition would be susceptible to momentary miscommunication. In contrast, if the modified instructions generally increased the salience of color contrast for the participants, then this manipulation should result in an overall increase in the production of RCAs across conditions, and not only in the atypical-color condition.

Method

Participants

Twenty-nine postgraduate students from the University of Groningen (Netherlands) took part in the experiment. Participants were all native speakers of Spanish (15 women) and participated for monetary compensation. The experiment was conducted at the University of Groningen because the author was collaborating in another project at the Psychology Department and the University of Groningen happens to have a large community of Spanish-speaking students. All participants had come to the Netherlands to complete their higher education. Ethics approval was obtained from the Psychology Department in Groningen.

Materials and Procedure

Experiment 2 used 4 × 4 grids in a within-participant design. Sixteen 4 × 4 grids were constructed, each including 12 clip-art objects and 4 empty cells, all randomly distributed in the grid space. Three types of objects were used as targets: animals, fruits and vegetables, and artifacts. These three target types were divided into three categories depending on the typicality of their color: stereotypical, variable and atypical color. Atypical colors never applied to the target category in real life (e.g., a pink banana or a blue camel). Each grid included four animals and four fruits and vegetables, two of each in stereotypical colors and two in atypical colors, plus four artifacts in variable colors. During the experiment participants had to refer to five items of each color type in a fixed random order. All the objects in the grids were different and so the use of color adjectives was redundant in all trials (see **Figure 3**). The first trial was treated as a warm-up and was not analyzed.

The grids of objects were presented on a computer monitor using E-Prime. Participants had to ask the experimenter to click on a specific object in each grid. For the participants' materials, fifteen 4 × 4 blank grids were printed on paper. A cross was placed in each blank grid in the cell that corresponded with the target object in the computer display. In order to facilitate synchronization between the experimenter and the participant, the grids of objects and the blank grids were numbered.

Participants had to ask the experimenter to click on the object in the display that corresponded with the cross on their blank grid. Participants sat behind the experimenter so that the experimenter could not see their paper grids but they could see the computer monitor in front of the experimenter. Participants were (falsely) told that the computer program randomized the objects in the grids and the experimenter was therefore naïve as to which object participants would ask for in each trial.

Given the simplicity of the task, participants were told that they were a control group in a developmental study investigating children's abilities to navigate two-dimensional spaces. Two types of instructions were used, Standard and Cautionary. In both instructions participants were told that they were going to play an interactive game in which they had to tell the experimenter to click on a specific figure in a grid of objects on the computer screen, using a cross on an empty paper grid to identify the target. The Cautionary instructions were identical to the Standard instructions with the exception of a paragraph at the end of the text in which participants were (falsely) told that in the pilot phase of the study, communication had sometimes broken down because participants did not pay enough attention to the objects in the grid and failed to notice that there were two objects of the

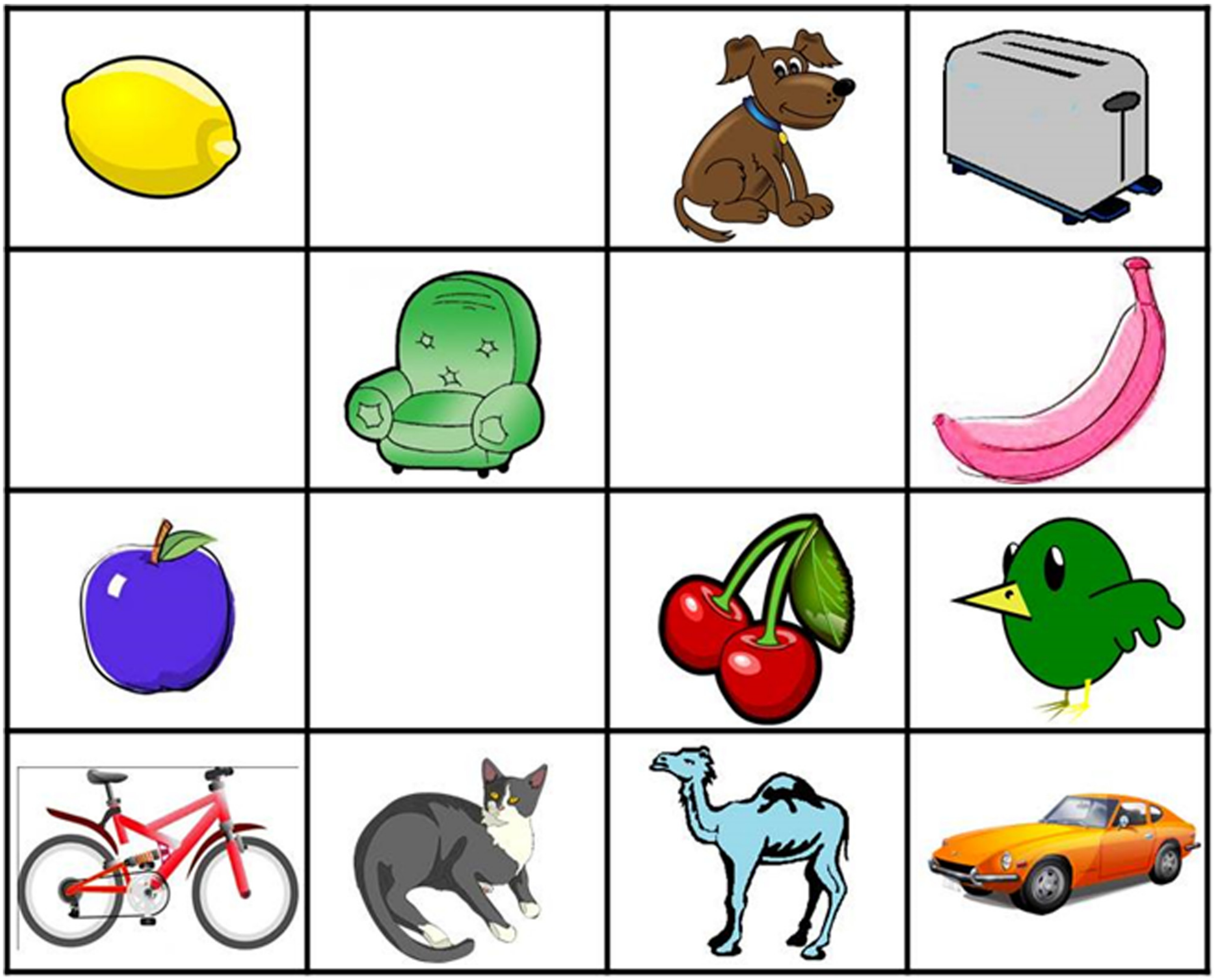


FIGURE 3 | Grid of objects from the Yellow Pig task. All grids included two fruits and vegetables of typical colors (yellow lemon and red cherries), two fruits and vegetables of atypical colors (pink banana and purple apple), two animals of typical colors (brown dog and gray cat), two animals of atypical colors (green bird and blue camel) and four artifacts of variable colors (red bike, green sofa, silver toaster, and orange car). In this particular trial the target was the pink banana.

same category. A false example was given in which a participant asked the experimenter to 'click on the box' and the experimenter had to ask which one: the big box or the small box. In reality, the grids of objects never included two objects of the same category, and so ambiguity was not an issue in any of the trials. Also, color was not mentioned in either type of instructions to avoid making it salient.

The whole task was administered in Spanish. The results of Experiment 1 suggest that English participants may have produced more RCAs than Spanish speakers. This, however, does not affect the predictions tested in Experiment 2 since the same pattern of results would be predicted for both groups of speakers across conditions (although the proportions of RCAs observed in each condition may have been higher for English speakers). Participants' requests were recorded and later coded by the experimenter as including or not including a RCA. Only referential expressions including both an adjective and a noun (e.g., 'El plátano rosa' – the pink banana) were coded as overspecific.

Results

Participants instructions conformed to the minimal or color-overspecific descriptions that were elicited (e.g., 'El plátano' or 'El plátano rosa' – the banana or the pink banana). The mean proportions of RCAs for each Color and Instruction type are shown in **Table 2**. Looking first at the effect of color typicality on the production of RCAs, a Friedman test revealed a significant difference among the three Color types across the two Instruction types, $X^2(2) = 28.2$, $p < 0.001$. *Post hoc* analyses of participants' RCA scores with the Wilcoxon test revealed that RCAs were produced significantly more often in the Atypical-color condition than in the Stereotypical-color condition, $Z = -3.47$, $p < 0.002$; also significantly more often in the Variable-color condition than in the Stereotypical-color

TABLE 2 | Mean proportions of redundant color adjectives (SD) produced in each color and instruction condition of the Yellow Pig task.

Instructions	Color		
	Stereotypical	Variable	Atypical
Standard	0 (0)	0.06 (0.15)	0.14 (0.36)
Cautionary	0 (0)	0.16 (0.17)	0.67 (0.38)

The task was conducted with Spanish speakers.

condition, $Z = -2.89$, $p < 0.005$; and significantly more often in the Atypical-color condition than in the Variable-color condition, $Z = -3.45$, $p < 0.002$.

Looking at the effect of Instruction type, as expected, no effect was observed on the production of RCAs in the Stereotypical-color condition, in which color terms were never used. The effect of Instruction type was non-significant in the Variable-color condition, $U = 68, Z = -1.92, p = 0.112$. In contrast, participants produced RCAs in the Atypical-color condition significantly more often after having received Cautionary instructions than after having received Standard instructions, $U = 36$, $Z = -3.24$, $p < 0.003$.

In addition, of the 14 participants in the Standard-instructions condition, 12 did not produce RCAs in any of the trials. Of the 15 participants in the Cautionary-instructions condition, only two did not use RCAs in any of the trials. A Chi-square test with continuity correction revealed that the difference between the number of participants who never produced RCAs (and those who sometimes did) in the two Instruction conditions is significant, $X^2(1,N = 29) = 12.4$, $p < 0.001$, with more participants never producing RCAs in the Standard-instruction condition than in the Cautionary-instructions condition.

Looking at item effects, the 10 participants who produced RCAs in the Variable-color condition (following either type of Instructions) always did so in the same one or two consecutive trials (i.e., a red pencil and/or a green cup). Moreover, 6 of these 10 instances were potential carry-over effects from a previous Atypical trial in which participants had overspecified color. This pattern of results suggests that the increase in the use of RCAs observed in the Variable-color condition between the Standard and the Cautionary instructions did not generalize over items (and was potentially related to the effect observed in the Atypical-color condition). By contrast, participants' use of RCAs in the Atypical-color condition was observed across all items in that category, thus revealing a more reliable effect of Instruction type.

Post-test

In order to rule out the possibility that the results of Experiment 2 reflected differences in the relative saliency of the target color in the different displays, grayscale saliency maps were created for the 15 slides employed in the study using Achanta et al. (2009)'s algorithm. The saliency maps were given to two naïve coders who ranked the 12 objects in each display according to their perceived salience (with white objects and black objects corresponding with the most and least salient objects in the display, respectively). Only the ranking of the target object was computed, with the highest ranking being adopted by default when there was a disagreement. The average ranking of the targets in the Variable-color condition was 6.6 (range: 8, 2, 5, 6, 12), in the Atypical-color condition was 5.6 (range: 2, 11, 2, 7, 6) and in the Stereotypical-color condition was 6.6 (range: 10, 4, 3, 12, 4). The results of this post-test using grayscale saliency maps suggest that the tendency to overspecify the color of atypical-color objects was not triggered by these targets being more perceptually salient than the targets in the Variable-color and Stereotypical-color conditions.

Discussion

The results of Experiment 2 confirm that color typicality has an effect on the production of RCAs in objects requests: while stereotypical colors were never used redundantly, variable colors were used redundantly significantly more often. This pattern of results replicates, with Spanish speakers, those reported by Sedivy (2003, 2004) with English speakers. As predicted, atypical colors were used redundantly significantly more often than variable and stereotypical colors. The difference between atypical and variable colors is particularly revealing, since the Variable-color condition was a neutral baseline for color typicality (i.e., the color of variable-color artifacts was neither stereotypical nor atypical). In line with the results reported by Westerbeek et al. (2015) with Dutch speakers, the results of Experiment 2 suggest that the less typical a color is for a given category token, the more likely it will be encoded in a request for the object.

The pattern of results observed in the Stereotypical and Atypical conditions with standard instructions is comparable to the results of Westerbeek et al. (2015). However, the proportion of RCAs for atypical color targets was much higher in Westerbeek et al. (2015) than in Experiment 2 (approximately 0.75 vs. 0.14, respectively). Leaving aside potentially important differences in the actual materials that were used in the two studies (which differed in type of objects and colors), a possible explanation for this difference is that Dutch speakers encode adjectives pre-nominally, while Spanish speaker do so post-nominally. The different results observed with these two groups of speakers therefore parallels the difference observed in Experiment 1 between English and Spanish speakers, indirectly supporting the hypothesis that adjective position is an important factor in the production of RCAs in face-to-face referential communication.

Regarding the issue of whether speakers are being cooperative when they mention atypical colors in object requests, the results of Experiment 2 suggest that participants may have tried to prevent the hearer from deriving the wrong presupposition and looking for a stereotypical target. Thus, those participants who received the cautionary instructions did not adopt a general strategy to describe the color of all types of targets in order to aid communication; instead, they did so mostly when referring to atypical-color objects, which were the only targets that could have caused momentary miscommunication. I interpret these results as evidence that RCAs can be cooperative in nature, although other factors and considerations may also be at play (e.g., the position of the adjective or the pertinence of color for the noun category, as suggested by the results of Experiment 1).

GENERAL DISCUSSION

Contrary to standard pragmatic models in the Gricean tradition, I have argued that speakers may be efficient when they produce RCAs in referential communication. If this is the case, speakers should produce RCAs in those situations in which they could reasonably expect that color would facilitate the hearer's visual search for the referent. An efficiency-based analysis of color overspecification is finer-grained than standard pragmatic analyses in terms of informativeness, and can therefore explain a number of perceptual factors that should affect a speaker's choice of referential expression (provided the speaker is rational and cooperative, as Gricean models assume).

As predicted, the production of RCAs in the present study was affected both by visual-contextual and semantic-category factors. Thus, speakers produced significantly more RCAs in polychrome displays than in monochrome displays, and did so more often when the adjective appeared in pre-nominal position (English) than in post-nominal position (Spanish). The results of Experiment 1 therefore suggest that speakers tend to produce RCAs when color may facilitate object identification for the hearer, hence behaving efficiently. This conclusion was also supported by the results of Experiment 2, where participants produced more RCAs when they were alerted to possible communication difficulties, suggesting that the use of RCAs can be cooperative in nature.

Semantic-category factors related to world knowledge also affected the production of RCAs, with English speakers producing twice as many RCAs when referring to clothes than to geometrical figures, and Spanish speakers producing four times as many (Rubio-Fernández, 2015). Moreover, English speakers produced significantly more RCAs when referring to clothes than when referring to geometrical figures in monochrome displays (Rubio-Fernández, under review). Finally, participants in Experiment 2 produced more RCAs for entities of atypical colors than for entities of variable and stereotypical colors, suggesting that our world knowledge affects our use of color adjectives; for example, when the color of an object violates our expectations (and hence those of our interlocutors).

The results of this study have implications for computational models of reference production, in particular Dale and Reiter's (1995) classic Incremental Algorithm, which incorporates Pechmann's (1989) finding that salient attributes such as color are sometimes overspecified. Because the Incremental Algorithm selects attributes in a preferred order, it is computationally simple and easy to implement (for a review of this and related algorithms, see Krahmer and van Deemter, 2012). However, from a psycholinguistic point of view, the Incremental Algorithm fails to incorporate the multiple factors that may affect the production of RCAs in referential communication. For example, the results of the present study show that in an otherwise identical situation, the use of color adjectives may vary depending on the syntactic position of the adjective (pre-nominal vs. post-nominal), the semantic category of the referent (e.g., comparable displays of clothes vs. geometrical figures), the typicality of the color of the referent (e.g., a yellow banana vs. a pink banana), and the speaker's disposition to maximize the chances of successful communication (see also Koolen et al., 2011).

Insensitive to all these sources of variation, the Incremental Algorithm produces RCAs because it treats color as a preferred attribute and never withdraws attributes once they have been selected (not even when the later inclusion of another attribute would render color redundant). However, because the algorithm checks the category of the object before its color, it never overspecifies color if the category is unique in the context (contrary to what was observed in this and other studies). Also, the Incremental Algorithm only overspecifies color if it has discriminatory value (e.g., it would never generate a color adjective in a monochrome display, contrary to what was observed in the Paper Dolls task). The results of this study therefore call for a more nuanced treatment of color in computational models of reference production, besides making it a preferred attribute that may be overspecified in very specific situations (for a discussion of various probabilistic revisions to the Incremental Algorithm, see Krahmer and van Deemter, 2012; van Deemter et al., 2012).

The results of this study also have implications for pragmatic models of reference production, which so far have failed to take perceptual factors into consideration. More specifically, I want to challenge the view that the use of RCAs is 'non-contrastive,' as opposed to those uses that are intended to establish a contrast between two objects of the same kind (Sedivy, 2004, 2007; Grodner and Sedivy, 2011). In my view, participants in this and earlier studies may have used RCAs in order to exploit a color contrast among different types of objects (e.g., Belke and Meyer, 2002; Sedivy, 2003; Koolen et al., 2013). Thus, participants may have asked the experimenter for 'the blue cup,' for example, in a situation where there was only one cup; however, if the cup was the only blue object in the display, then color would have been used contrastively. This interpretation of the effect of color contrast on the production of color adjectives calls for a revision of the pragmatic notion of *referential contrast*.

The canonical function of an adjective in an object request is to exploit a contrast between the intended referent and other objects of the same kind, which would allow the interlocutor to uniquely identify the target object against its competitors (e.g., a plastic cup vs. a paper cup). Contrary to the standard view, I want to propose that in the case of prenominal color adjectives, referential contrast may be established across categories, rather than within a given category (the way it is established for material and relative adjectives; Sedivy, 2003, 2004). According to this definition, prenominal color adjectives are used contrastively whenever there is a color contrast in the visual context that the speaker could exploit for efficient referential communication (e.g., a blue cup vs. a red jug). Those situations where prenominal color adjectives are used contrastively in the canonical sense of the term to distinguish between two objects of the same kind (e.g., a blue cup vs. a red cup) are merely a special case of color-contrastive uses.

DOES COLOR OVERSPECIFICATION POSE A CHALLENGE TO GRICEAN PRAGMATICS?

According to Grice's (1975) Maxim of Quantity, speakers should try to provide their interlocutors with as much information as they need, but not more. The extent to which overspecification poses a challenge to Gricean models of referential communication has been a recurrent theme in the pragmatics literature (e.g., Sedivy, 2003, 2004, 2007; Engelhardt et al., 2006; Grodner and Sedivy, 2011; Heller et al., 2012). The most extreme position in this debate has been adopted by Engelhardt et al. (2011), who went so far as to argue that overspecific referential expressions 'impair comprehension' (but cf. Sonnenschein and Whitehurst, 1982; Mangold and Pobel, 1988; Maes et al., 2004; Davies and Katsos, 2010; Arts et al., 2011b). The results of the study by Engelhardt et al. (2011) are not entirely surprising, however, since these authors investigated the effect of size and color overspecification using minimal displays of two figures in which the hearer had to identify a target following a modified description (e.g., 'the red square' or 'the big star'). Given the simplicity of the hearer's visual search, it is only to be expected that color would not facilitate object identification when the display included two different figures (e.g., a red square and a blue circle vs. a red square and a blue square).

Two patterns of results seem to support this interpretation of the results of Engelhardt et al. (2011): eye-tracking studies of adjectival modification (e.g., Sedivy et al., 1999; Sedivy, 2003, 2004; Rubio-Fernández, under review) have repeatedly shown that listeners are able to visually identify a referent as soon as they have enough information to do so, even when the adjective is redundant. In Engelhardt et al.'s (2011) study, size and color were distinctive properties of the target in both the redundant and the contrastive conditions, which means participants should have been able to visually identify the target referent in hearing the adjective. However, Engelhardt et al. (2011) did not use eye tracking to measure reference resolution during processing; instead, they asked participants to press a right/left key to indicate the position of the target on the screen. The longer response times observed in the redundant condition suggest that participants' responses did not measure visual identification alone (which should have been comparable in both conditions), but also reflected an implicit pragmatic judgment by comparison to the contrastive condition.

One reason why participants may have found the overspecific descriptions in Engelhardt et al. (2011) pragmatically infelicitous is because they were unnatural in the visual context: Rubio-Fernández (under review) observed that speakers never overspecified the size of a target in a 2-figure display, and did so less than 25% of the time with color adjectives. When using larger displays, however, both size and color were overspecified over 60% of the time. That the overspecified descriptions used by Engelhardt et al. (2011) were highly unnatural might also explain the early N400 observed in that condition, which the authors interpreted as reflecting either a semantic integration problem or low predictability.

On reflection, what is more remarkable in the study by Engelhardt et al. (2011) is their interpretation of their results as in line with Grice's (1975) model of verbal communication. After all, Grice's model rests not only on the Maxim of Quantity, but most importantly on the Cooperative Principle and the general assumption that speakers and hearers interact as rational agents. Therefore, if redundant referential expressions impair communication, as Engelhardt et al. (2011) claim, why do speakers overspecify their referential expressions as often as they do? One would assume that a rational and cooperative speaker who had a choice between referring to 'the t-shirt' or 'the yellow t-shirt' would not choose (systematically, sometimes) the modified description if that would impair the hearer's comprehension.

It seems safe to assume that speakers are being rational and cooperative when they produce RCAs that could facilitate the interlocutor's search for the referent (e.g., ask for 'the blue cup' in a situation where there is only one cup, but it is also the only blue object in a relatively dense display). But what about those RCAs that are produced in monochrome displays? In the present study only English speakers produced such RCAs when referring to clothes, while Spanish speakers did not. Moreover, Rubio-Fernández (under review) reports that English speakers produced zero rates of color overspecification in monochrome displays of geometrical figures using the same task. These results suggest that the tendency to use RCAs in monochrome displays observed in Experiment 1 was driven by the pertinence of color for clothes and the general tendency of English speakers to overspecify color.

However, using RCAs to refer to clothes cannot be considered as irrational or un-cooperative behavior since the pertinence of color for clothes may be so high that hearers expect that the color of clothes be encoded in referential communication. Along these lines, Dale and Reiter (1995) have argued that one reason why speakers may sometimes use RCAs when color has no discriminatory power in the context is because they are using reference scripts that determine which attributes are expected for a certain semantic category (for a related view using 'default descriptions,' see Sedivy, 2003, 2004, 2007; Grodner and Sedivy, 2011). This could be the case for the color of clothes and shoes in the English language, as suggested by collocations such as 'black tie', 'little black dress,' 'the red shoes' or 'white collar workers.'

CONCLUSION

Those color adjectives that are not necessary to establish unique reference are traditionally considered redundant or over-informative, even though they may be efficient (insofar as they may facilitate the interlocutor's search for the object) and/or pertinent for the requested object (insofar as color is important for the semantic category). Therefore, traditional pragmatic analyses cast in terms of informativeness alone fall short of explaining the ubiquitous use of RCAs in referential communication and the kind of factors that affect these uses. An analysis in terms of efficiency and pertinence, however, reveals that the use of RCAs is in line with Gricean pragmatics.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

ACKNOWLEDGMENTS

The research reported in this study was supported by a Euro-XPrag Grant from the European Science Foundation (*The cognitive costs and benefits of using color adjectives* – Ref. No. 3844) and by a Young Research Talent Grant from the Research Council of Norway (*Saying enough without saying too much* – Project No. 230718). Open Access publication costs were granted by the University of Oslo (Ref. No. 855953). All three grants are gratefully acknowledged. The author would also like to thank Chris Cummins, Bart Geurts, Sam Glucksberg, and Susanne Grassmann for many colorful discussions on the pragmatics of color adjectives.

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Süsstrunk, S. (2009). "Frequency-tuned salient region detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2009* (Miami, FL: IEEE), 1597–1604.

Arnold, J. (2008). Reference production: production-internal and addressee-oriented processes. *Lang. Cogn. Process.* 23, 495–527. doi: 10.1080/01690960801920099

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011a). Overspecification in written instruction. *Linguistics* 49, 555–574. doi: 10.1515/ling.2011.017

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011b). Overspecification facilitates object identification. *J. Pragmat.* 43, 361–374. doi: 10.1016/j.pragma.2010.07.013

Belke, E. (2006). Visual determinants of preferred adjective order. *Vis. Cogn.* 14, 261–294. doi: 10.1080/13506280500260484

Belke, E., and Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: analyses of viewing patterns and processing times during "same"-"different" decisions. *Eur. J. Cogn. Psychol.* 14, 237–266. doi: 10.1080/09541440143000050

Brennan, S. E., Galati, A., and Kuhlen, A. K. (2010). Two minds, one dialog: coordinating speaking and understanding. *Psychol. Learn. Motiv.* 53, 301–344. doi: 10.1016/S0079-7421(10)53008-1

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., and Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *J. Mem. Lang.* 47, 30–49. doi: 10.1006/jmla.2001.2832

Chambers, C. G., Tanenhaus, M. K., and Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *J. Exp. Psychol. Learn. Mem. Cogn.* 30, 687–696.

Clark, H. H., and Marshall, C. R. (1981). "Definite reference and mutual knowledge," in *Elements of Discourse Understanding*, eds A. K. Joshi, B. Webber, and I. A. Sag (Cambridge: Cambridge University Press), 10–63.

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Clarke, A. D. F., Elsner, M., and Rohde, H. (2013). Where's wally: the influence of visual salience on referring expression generation. *Front. Percept. Sci.* 4:329. doi: 10.3389/fpsyg.2013.00329

Clarke, A. D. F., Elsner, M., and Rohde, H. (2015). Giving good directions: order of mention reflects visual salience. *Front. Psychol.* 6:1793. doi: 10.3389/fpsyg.2015.01793

Dale, R., and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902_3

Davies, C., and Katsos, N. (2010). Over-informative children: production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua* 120, 1956–1972. doi: 10.1016/j.cognition.2011.02.015

Engelhardt, P. E., Bailey, K. G. D., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Engelhardt, P. E., Demiral, Ş. B., and Ferreira, F. (2011). Overspecified referring expressions impair comprehension: an ERP study. *Brain Cogn.* 77, 304–314. doi: 10.1016/j.bandc.2011.07.004

Galati, A., and Brennan, S. E. (2010). Attenuating information in spoken communication: for the speaker, or for the addressee? *J. Mem. Lang.* 62, 35–51. doi: 10.1016/j.jml.2009.09.002

Geurts, B. (2010). *Quantity Implicatures*. Cambridge: Cambridge University Press.

Grice, H. P. (1975). "Logic and conversation," in *Speech Acts*, eds P. Cole and J. Morgan (New York, NY: Academic Press), 41–58.

Grodner, D., and Sedivy, J. (2011). "The effects of speaker-specific information on pragmatic inferences," in *The Processing and Acquisition of Reference*, eds N. Pearlmutter and E. Gibson (Cambridge, MA: MIT Press), 239–272.

Hanna, J. E., and Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *J. Mem. Lang.* 57, 596–615. doi: 10.1016/j.jml.2007.01.008

Hanna, J. E., and Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cogn. Sci.* 28, 105–115. doi: 10.1207/s15516709cog2801_5

Heller, D., Gorman, K. S., and Tanenhaus, M. K. (2012). To name or to describe: shared knowledge affects referential form. *Top. Cogn. Sci.* 4, 290–305. doi: 10.1111/j.1756-8765.2012.01182.x

Huettig, F., and Altmann, G. T. (2011). Looking at anything that is green when hearing "frog": how object surface color and stored object color knowledge influence language-mediated overt attention. *Q. J. Exp. Psychol.* 64, 122–145. doi: 10.1080/17470218.2010.481474

Keysar, B. (1997). Unconfounding common ground. *Discourse Process.* 24, 253–270. doi: 10.1080/01638539709545015

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Koolen, R., Goudbeek, M., and Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cogn. Sci.* 37, 395–411. doi: 10.1111/cogs.12019

Krahmer, E., and van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI_a_00088

Lockridge, C. B., and Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychon. Bull. Rev.* 9, 550–557. doi: 10.3758/BF03196312

Maes, A., Arts, A., and Noordman, L. (2004). Reference management in instructive discourse. *Discourse Process.* 37, 117–144. doi: 10.1207/s15326950dp3702_3

Mangold, R., and Pobel, R. (1988). Informativeness and instrumentality in referential communication. *J. Lang. Soc. Psychol.* 7, 181–191. doi: 10.1177/0261927X8800700403

Mitchell, M., Reiter, E., and Deemter, K. V. (2013). "Typicality and object reference," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (CogSci)*, Berlin.

Nadig, A., and Sedivy, J. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychol. Sci.* 13, 329–336. doi: 10.1111/j.0956-7976.2002.00460.x

Neider, M. B., Chen, X., Dickinson, C. A., Brennan, S. E., and Zelinsky, G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychon. Bull. Rev.* 17, 718–724. doi: 10.3758/PBR.17.5.718

Paraboni, I., van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: making referents easy to identify. *Comput. Linguist.* 33, 229–254. doi: 10.1162/coli.2007.33.2.229

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 89–110. doi: 10.1515/ling.1989.27.1.89

Pogue, A., Kurumada, C., and Tanenhaus, M. K. (2015). Talker-specific generalization of pragmatic inferences based on under– and over-informative prenominal adjective use. *Front. Psychol.* 6:2035. doi: 10.3389/fpsyg.2015.02035

Reali, F., Spivey, M. J., Tyler, M. J., and Terranova, J. (2006). Inefficient conjunction search made efficient by concurrent spoken delivery of target identity. *Percept. Psychophys.* 68, 959–974. doi: 10.3758/BF03193358

Reiter, E. (1991). A new model of lexical choice for nouns. *Comput. Intell.* 7, 240–251. doi: 10.1111/j.1467-8640.1991.tb00398.x

Richardson, D. C., and Dale, R. (2005). Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cogn. Sci.* 29, 1045–1060. doi: 10.1207/s15516709cog0000_29

Rubio-Fernández, P. (2015). "Redundancy is efficient – and effective, too," in *Paper Presented at the XI Conference on Architectures and Mechanisms for Language Processing (AMLaP) 2015* (Malta: University of Malta).

Sedivy, J. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguist. Res.* 32, 3–23. doi: 10.1023/A:1021928914454

Sedivy, J. (2004). "Evaluating explanations for referential context effects: evidence for Gricean mechanisms in online language interpretation," in *Approaches to Studying World-Situated Language Use*, eds J. Trueswell and M. Tanenhaus (Cambridge, MA: MIT Press), 345–364.

Sedivy, J. (2007). Implicatures in real-time conversation: a view from language processing research. *Philos. Compass* 2, 475–496. doi: 10.1111/j.1747-9991.2007.00082.x

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition* 71, 109–147. doi: 10.1016/S0010-0277(99)00025-6

Sonnenschein, S., and Whitehurst, G. J. (1982). The effects of redundant communications on the behavior of listeners: does a picture need a thousand words? *J. Psycholinguist. Res.* 11, 115–125. doi: 10.1007/BF01068215

Spivey, M. J., Tyler, M. J., Eberhard, K. M., and Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychol. Sci.* 12, 282–286. doi: 10.1111/1467-9280.00352

Tarenskeen, S., Broersma, M., and Geurts, B. (2015). Overspecification of color, pattern, and size: salience, absoluteness, and consistency. *Front. Psychol.* 6:1703. doi: 10.3389/fpsyg.2015.01703

van Deemter, K., Gatt, A., van Gompel, R. P., and Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Top. Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01187.x

van der Sluis, I., and Krahmer, E. (2007). Generating multimodal references. *Discourse Process.* 44, 145–174. doi: 10.1016/j.ejrad.2011.08.005

Westerbeek, H., Koolen, R., and Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Front. Psychol.* 6:935. doi: 10.3389/fpsyg.2015.00935

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer RK and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 November 2015
doi: 10.3389/fpsyg.2015.01703

Overspecification of color, pattern, and size: salience, absoluteness, and consistency

Sammie Tarenskeen[1]*, *Mirjam Broersma*[2,3] *and Bart Geurts*[1]

[1] *Department of Philosophy, Radboud University, Nijmegen, Netherlands,* [2] *Centre for Language Studies, Radboud University, Nijmegen, Netherlands,* [3] *Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands*

The rates of overspecification of color, pattern, and size are compared, to investigate how salience and absoluteness contribute to the production of overspecification. Color and pattern are absolute and salient attributes, whereas size is relative and less salient. Additionally, a tendency toward consistent responses is assessed. Using a within-participants design, we find similar rates of color and pattern overspecification, which are both higher than the rate of size overspecification. Using a between-participants design, however, we find similar rates of pattern and size overspecification, which are both lower than the rate of color overspecification. This indicates that although many speakers are more likely to include color than pattern (probably because color is more salient), they may also treat pattern like color due to a tendency toward consistency. We find no increase in size overspecification when the salience of size is increased, suggesting that speakers are more likely to include absolute than relative attributes. However, we do find an increase in size overspecification when mentioning the attributes is triggered, which again shows that speakers tend to refer in a consistent manner, and that there are circumstances in which even size overspecification is frequently produced.

Keywords: referential overspecification, attribute selection, color, salience, absoluteness, consistent responses

OPEN ACCESS

Edited by:
Emiel Krahmer,
Tilburg University, Netherlands

Reviewed by:
Paul Edward Engelhardt,
University of East Anglia, UK
Ivandre Paraboni,
University of Sao Paulo, Brazil

****Correspondence:***
Sammie Tarenskeen
s.tarenskeen@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *18 July 2015*
Accepted: *22 October 2015*
Published: *06 November 2015*

Citation:
Tarenskeen S, Broersma M and Geurts B (2015) Overspecification of color, pattern, and size: salience, absoluteness, and consistency. Front. Psychol. 6:1703. doi: 10.3389/fpsyg.2015.01703

1. INTRODUCTION

When speakers refer to objects, they do not always limit themselves to giving information that is strictly necessary for the addressee to identify the referent. In other words, they sometimes produce *overspecification* instead of *minimal specification* (e.g., Pechmann, 1989; Engelhardt et al., 2006; Arts et al., 2011b). Imagine, for example, a speaker requesting her addressee to pass her a yellow cup, which happens to be surrounded by blue plates and bowls. Although the speaker need not include a color adjective to enable her addressee to identify the referent, because there is only one cup present, experimental work suggests that she would be more likely to utter (1-b) than (1-a) in this situation, and hence, to produce *color* overspecification.

(1) a. Please pass me the cup.
b. Please pass me the *yellow* cup.

Experimental findings suggest that there is something special about color in reference: including color is preferred over including various other attributes, most notably size. When it is necessary to include either color or size to get a unique description of the referent, color is more often included than size (Belke and Meyer, 2002). Color is also more likely to be included redundantly than size:

for example, when referring to a *small* yellow cup surrounded by *big* cups in yellow, red, and green, many speakers will not only select size, which is both necessary and sufficient for identification of the referent, but also color, which is neither necessary nor sufficient (Pechmann, 1989). When referring to an object that is unique in its type, as in the situation above, speakers often include color as well (Koolen et al., 2013), even though no modification (e.g., an adjective) is needed at all in that case. Most extremely, even when all objects in the visual context have the same color as the referent, color is sometimes mentioned (Mangold and Pobel, 1988; Belke and Meyer, 2002; Koolen et al., 2015).

In this paper, we investigate the seemingly special status of color in reference production, and in overspecification in particular. We do this by comparing color with two other attributes: pattern and size. Whereas color and size overspecification have been investigated before, the study of reference to pattern is virtually unexplored. Pattern is an interesting attribute because it is like color—but unlike size—in being both *salient* and *absolute*. As these two factors have been suggested to explain why speakers produce color overspecification, comparing the three attributes will enable us to systematically tease apart, for the first time, the effect of the two factors on the tendencies to include different attributes redundantly.

We present a series of four language production experiments. In our first experiment, we compare the rates of color overspecification with the corresponding rates of pattern and size overspecification. In one follow-up experiment, we then assess the effect of salience and absoluteness. In two other follow-up experiments, we assess the effect of *consistency*, that is, the tendency to reuse previous expressions and constructions, by varying color, pattern, and size both within and between participants, and by triggering selection of the three attributes.

2. SALIENCE, ABSOLUTENESS, AND CONSISTENCY

In this section, we discuss the literature on referential overspecification. In Section 2.1, we introduce the notion of salience as an important factor in attribute selection. The role of salience and absoluteness in the preference that speakers appear to have for including color is elaborated on in Section 2.2. In Section 2.3, we discuss experimental work on the speakers' tendency to behave consistently. Finally, we introduce the series of experiments that we conducted in more detail in Section 2.4.

2.1. Salience and Overspecification

A question in the research of referring expressions production that has received much attention lately is how speakers select attributes when producing definite descriptions (for a recent overview, see van Deemter et al., 2012). A factor that is currently thought to be central to attribute selection is *salience* (e.g., Gatt, 2007; Arts et al., 2011a; Koolen et al., 2011). An object's attribute can be salient for various reasons, and is then more likely to be selected by a speaker who intends to refer to this object. This may result in overspecification, as salient attributes are not always necessary to enable the addressee to identify the referent.

The basic idea of selecting salient attributes is intuitive: speakers tend to select the attributes according to the degree to which their attention is attracted by them. In the literature on salience and visual perception, visual or perceptual salience is considered to be a property of *objects*, which may be defined in terms of surprise (Itti and Baldi, 2009). Surprise can occur on a low level, for example, when an object is unique on one or more dimensions (Treisman and Gelade, 1980), such as a blue round candy among red cubic candies. It can also occur on a higher level, induced by world knowledge (Franke, 2012): a blue banana will in general be more salient than a yellow banana.

In the literature on reference production, it is assumed (often implicitly) that not only objects, but also *attributes* of objects vary in salience (e.g., Davies and Katsos, 2013). Attributes that are unique in a given context, like color and shape in the candies example above, may be salient, and attributes that are surprising due to world knowledge, such as the color of a blue banana, may be salient as well, analogously to factors that determine the salience of objects. Indeed, speakers tend not to include redundant color adjectives when referring to objects strongly associated with a specific color, for instance, the color of a yellow banana (Sedivy, 2003), which is entirely as expected and therefore not particularly salient. If a referent has an unexpected color, however, color overspecification is much more likely to occur (Westerbeek et al., 2014). Davies and Katsos (2013) show that speakers are more likely to produce overspecification when objects have salient attributes than when they do not.

It seems a good idea to select attributes that are salient, not only because it is easy for the speaker, as has often been suggested (Mangold and Pobel, 1988; Davies and Katsos, 2013; Koolen et al., 2013), but also, and perhaps more importantly, from a communicative point of view (cf. Arts et al., 2011b; Koolen et al., 2011; Davies and Katsos, 2013). If an attribute attracts the speaker's attention, it is likely that it will attract the attention of her addressee as well, which probably increases the likelihood that it is useful in the process of identifying the referent. Not all salient attributes are necessary for referent identification, however, and selecting them may therefore result in overspecification. Although the word "overspecification" may have a negative flavor, suggesting that the expression is *too* specific, overspecification need not be cumbersome and may even be beneficial, as the benefits of a strictly redundant but salient attribute in the comprehension process may often outweigh the risk that the addressee is hindered by its redundancy. Indeed, there is evidence that overspecification can speed up the process of referent identification (Sonnenschein and Whitehurst, 1982; Mangold and Pobel, 1988; Paraboni et al., 2007; Arts et al., 2011b; but see Engelhardt et al., 2006, 2011). An eyetracking study on the processing of size and color adjectives suggests that redundant size adjectives may be confusing for addressees, whereas redundant color adjectives are not (Sedivy et al., 1999). Another study on the comprehension of overspecified expressions suggests, moreover, that non-salient redundant attributes are more likely to hinder the addressee than salient redundant attributes (Davies and Katsos, 2013).

In sum, there seems to be a tendency to select salient attributes, even if this results in overspecification. Redundancy can hinder the comprehension process, but as salient attributes are likely to be helpful in referent identification, including a redundant but salient attribute may often be beneficial.

2.2. The Color Preference

The literature suggests that speakers tend to include color more often than other attributes, and that color overspecification is more common than overspecification of other attributes. Two features of color have been argued to contribute to this preference: salience and absoluteness. We will discuss both features in this section. An overview of salience and absoluteness of color, pattern, and size is presented in **Table 1**.

2.2.1. Salience

In line with the view that speakers tend to select salient attributes, it has been argued that color is preferred because it is *intrinsically* salient (Arts et al., 2011a; Gatt et al., 2013; Koolen et al., 2013). The common view is that intrinsically salient attributes are noticed immediately, and before other attributes: they are "perceived earlier" (Gatt, 2007) and "immediately grab [the speakers'] attention" (Koolen et al., 2013). It has also been suggested that color is more likely to "pop out" than other attributes (Westerbeek et al., 2014): intuitively, one green candy in a jar surrounded by red ones is more likely to be noticed than one small candy surrounded by big ones, or one cubic candy surrounded by round ones.

Indeed, color is one of the features computed in the earliest stages of human visual processing (Livingstone and Hubel, 1988), and can be considered a primary cue in visual perception. It has been found that objects in a color that is contextually unique can grab the attention in visual search, even if color is irrelevant to the task (Theeuwes, 1992; Turatto and Galfano, 2001). Color also tends to be more helpful in visual search than other attributes, such as size and shape (Williams, 1966; Christ, 1975). Color contrast between items thus seems to be an extremely powerful cue in visual perception. In this respect, color may be different from other visual attributes, and also from non-visual attributes, like material, some of which have been found to be included redundantly less often than color (see Mangold and Pobel, 1988, for shape, Arts et al., 2011b, for size, and Sedivy, 2005, for size and material).

When examining experimental stimuli from previous experiments, however, we observed that colors in experimental stimuli tend to be bright and/or highly contrastive, while differences in size are usually rather modest (e.g., Arts et al., 2011b; Koolen et al., 2011). We argue, then, that previous findings do not necessarily show that color is preferred over size due to a difference in salience. Rather, the specific colors and color contrasts used in those experiments may have been more salient than the size contrasts used, resulting in higher rates of color overspecification. Recently, the preference for color over size was found to disappear when the size contrast between the referent and other objects was increased (van Gompel et al., 2014). Along the same lines, speakers may be less inclined to produce color overspecification when the color contrast is low or when colors are not particularly vivid than when colors are bright and contrastive (Tarenskeen et al., in preparation). In sum, it is not evident that, for example, a pale blue candy surrounded by mint green ones is more likely to get the attention than a huge candy surrounded by tiny ones.

In the study conducted by van Gompel et al. (2014), *competition* between color and size was investigated. In the condition relevant for our study, the referent was different from the other objects in the array in color and size but not in type. For example, the referent was a small red candle and the other objects were a big blue and a big black candle. When the size contrast was low, participants included color but not size in 79% of the cases, and size but not color in only 2% of the cases. When the contrast was high, however, color but not size was included in only 27% of the cases, while the rate of referring expressions including size but not color increased to 23%. Importantly, it was always necessary to include either color or size. Hence, overspecification occurred only when *both* color and size were included. This set-up is suitable for studying attribute preferences, but not for comparing attributes with respect to how likely they are to be added redundantly, which is the aim of the present study. To be able to compare the rates of color, pattern, and size overspecification, we present participants with arrays in which the referent is unique in its type (for example, if the referent is a dress, none of the other objects in the array is a dress). Thus, adding an extra attribute always results in overspecification. As van Gompel et al. (2014), we manipulate the size contrast between the referent and surrounding objects. While they investigate the effect of size contrast on the choice for including size vs. color, we assess the effect of size contrast on the production of size overspecification.

While we vary the salience of size, we keep the two other attributes constant in being high in salience. Unlike color and size, pattern is virtually unexplored in the literature on reference production. In the only study investigating pattern in reference production, Gatt et al. (2013) found that speakers prefer color over both pattern and size. As in van Gompel et al.'s study, however, they investigated competition between attributes, using arrays in which the referent was not unique in its type. Moreover, they used a single superimposed shape (a circle, a diamond, or a square) on a brightly colored picture as patterns, e.g., a green bottle with a circle-shaped patch on it. Such patterns are probably not very salient, and pictures with one little figure would not normally be called "patterned". The use of striking colors may have decreased the salience of pattern even more. This thus leaves the crucial question open whether speakers are also more likely to produce color than pattern overspecification in a situation where pictures have salient patterns but no other salient attributes. The present study aims to address this question by depicting patterned objects which are completely striped or spotted and do not have any other striking attributes. If color overspecification is produced frequently because of its intrinsic salience, a high rate of pattern overspecification is expected too, as pattern may be highly salient as well. On the other hand, a high rate of size overspecification is only expected if size is made salient. In Section 2.4, we elaborate on this further.

2.2.2. Absoluteness

According to Pechmann (1989) and Belke and Meyer (2002), speakers tend to select color before size because color is an absolute attribute, whereas size is relative[1]. That is, a speaker need not take into account objects surrounding the referent in order to determine its color[2], while she normally has to do this to determine whether the referent is big or small. Pechmann points out that as speech is produced incrementally, the speaker can start to articulate the referent's color while examining the context in order to find out which additional attributes are required for a unique description, which may result in color overspecification. Pechmann's argument is in line with eyetracking results which indicate that speakers often start producing color adjectives before fixating on an item of the same type but a different color in the array (e.g., a blue cup when the referent is a yellow cup), while they rarely start producing size adjectives before fixating on a size-contrastive item (Brown-Schmidt and Konopka, 2011).

Two findings indicate that absoluteness alone does not explain the color preference. First, not all absolute attributes tend to be redundantly included in referring expressions. Although shape is an absolute attribute, shape overspecification has been found to occur less frequently than color overspecification (Mangold and Pobel, 1988; Arts et al., 2011b). In another study, material, which is also an absolute attribute, was included redundantly as infrequently as size, even though size is a relative attribute (Sedivy, 2005).

The second indication that absoluteness alone does not explain the color preference is that size adjectives usually precede both redundant and non-redundant color modifiers (e.g., "the big red car," Sproat and Shih, 1991; Cinque, 1994), while according to Pechmann's account, redundant color modifiers should in general precede size modifiers ("the red big car"). After all, color overspecification is due to speakers starting their referring expression after selecting color but before selecting size. In Pechmann's production study, speakers of Dutch indeed produced color before size adjectives sometimes, even though they would normally prefer the reverse order (Sproat and Shih, 1991, p. 580). However, in two studies with speakers of German and English, who have the same adjective order preference as speakers of Dutch (Cinque, 1994), color overspecification was produced frequently, but color hardly ever preceded size (Belke, 2006). This indicates that color overspecification is often not due to articulating color adjectives before selecting size, as Pechmann proposes. It is possible, however, that color is normally *selected* before size, without necessarily being *articulated* before selecting size (see also Belke and Meyer, 2002).

Although the distinction between absolute and relative attributes thus cannot entirely explain the asymmetry between color and size, the fact that color is absolute while size is relative is likely to play a role in the preference for color over size in reference. In the present study, we take into account the role of absoluteness by comparing color both to size, which is relative, and to pattern, which is absolute.

2.3. Consistency

Our main interest in this paper is in the overspecification of three different attributes that vary in salience and in being absolute or relative: color, pattern, and size. Additionally, we investigate the way in which the rates of overspecification of the three attributes may affect one another. Experimental studies show that speakers have a preference for sticking to previously used expressions and constructions (e.g., Brennan and Clark, 1996; Pickering and Garrod, 2004; Goudbeek and Krahmer, 2012). In this paper, we investigate the relation between this preference and tendencies to include one attribute but not another one. For example, if speakers have a preference for including color but not including size, a preference for consistency may result in a decrease in the rate of color overspecification, or an increase in the rate of size overspecification.

Recently, the attention of some researchers has been attracted by the high amount of variation *across* speakers when producing referring expressions in experimental settings. It was found that machine learning models predict human-produced referring expressions better when they take into account both speaker identity and characteristics of the visual context than when they only use visual characteristics (Viethen and Dale, 2010; see also Mitchell et al., 2011; Ferreira and Paraboni, 2014). Since machine learning models that used speaker identity based their predictions on previously produced referring expressions, this finding suggests not only that speakers strongly differ in their referring behavior, but also that individual speakers tend to be consistent in the way they refer. Indeed, a basic assumption in psychological research is that variation between participants is higher than variation within participants, which is why participants are often modeled as random variables in statistic analyses (e.g., Baayen et al., 2008).

The finding that speakers tend to refer in a consistent way is reminiscent of the well-established tendency to reuse referring expressions that have been used earlier in the conversation by one of the interlocutors. For example, Brennan and Clark (1996) showed that speakers who use a specific term instead of a basic-level term in order to avoid ambiguity, such as "the loafer" in a context with several kinds of shoes, tend to stick to this term even in contexts where the basic-level term would not lead to ambiguity any longer, such as 'the loafer' in a context where the loafer is the only shoe. Analogously, speakers were found to reuse constructions for the same referents by including modifiers that were redundant in the current context but necessary in preceding contexts (Van Der Wege, 2009).

More generally, speakers can be primed to include attributes that would normally be dispreferred, such as the orientation of the referent where its color would have been sufficient, too (Goudbeek and Krahmer, 2012). Another study suggests that attribute selection is affected by the linguistic context more than by some visual factors that are often expected to be influential, such as the degree to which the referent's attributes are unique in

[1] Size is usually considered to be a relative attribute because in experimental studies of reference, speakers refer to size by using gradable adjectives like "big" and "small," and not absolute measures such as centimeters.

[2] This is not strictly speaking true, as color perception is in fact highly sensitive to various features of the visual context. However, colors used for experimental stimuli are almost always bright, saturated colors that are highly typical for the color categories they fall into, being minimally sensitive to the context, rendering color practically an absolute attribute.

the visual context, called discriminatory power[3] (Viethen et al., 2014). They found that learning models of reference production that take into account features of previously produced referring expressions predicted human-produced expressions better than models selecting attributes based on discriminatory power, which is also in line with Gatt et al. (2013). The tendency to reuse words in experimental settings has been found outside the realm of reference as well (see e.g., Alferink and Gullberg, 2014).

In our study, we investigate whether due to a tendency toward consistency, the tendencies to include one attribute but not another can affect one another. We also assess whether, in line with Goudbeek and Krahmer (2012), mentioning the three attributes can trigger even size overspecification, which is normally produced infrequently. Our study is not intended, however, to assess the mechanisms that underpin consistency in reference production. Currently, a debate is going on about those mechanisms. One position is that in dialogue, interlocutors establish *conceptual pacts* (Brennan and Clark, 1996): they reuse referring expressions when talking to the same partner and expect their partner to do the same. This view presupposes that interlocutors keep track of their common ground, that is, the information that is mutually shared between them. According to the alternative account, interlocutors automatically *align* their representations on all linguistic levels (Pickering and Garrod, 2004). The central claim is that interlocutors do not need to keep track of their common ground, memory processes like priming normally being sufficient for proper alignment. That is, interlocutors reuse referring expressions because those expressions are salient due to their being primed by their previous usages. It is uncontroversial that priming is a mechanism present in both language production and comprehension: there is substantial evidence for semantic priming (e.g., Meyer and Schvaneveldt, 1971; Neely, 1976), phonological priming (e.g., Bock, 1986a; Grainger and Ferrand, 1996), and syntactic priming (e.g., Bock, 1986b; Potter and Lombardi, 1998). What researchers in the present debate essentially disagree about, however, is whether interlocutors routinely take into account their common ground when producing and comprehending utterances in a way that goes beyond automatic priming mechanisms (see amongst many others, Brown and Dell, 1978; Lockridge and Brennan, 2002; Pickering and Garrod, 2004; Yoon and Brown-Schmidt, 2014).

In sum, speakers often reuse words and constructions that were used earlier in the discourse, having a preference for consistency. They tend to do this even if there is in fact a good reason to switch to a different construction, like the changed context in Brennan and Clark's (1996) experiment, or the general preference for other attributes than orientation, as in Goudbeek and Krahmer's (2012) experiment. Consistency in reference production may be due to considerations of the interlocutors' common ground or to simple priming mechanisms. However, we are neutral as to what mechanisms may result in the effects we find, although we will discuss some possibilities in Section 7.

2.4. The Present Study

The present study investigates, in the first place, tendencies to include various attributes in referring expressions, even if this results in overspecification, and the way in which salience and absoluteness contribute to these tendencies. In order to do this, we conduct four language production experiments in which speakers use referring expressions to refer to pictures of objects that vary in color, pattern, and size. We compare the proportions of overspecification of the three attributes. Our study is the first to compare attributes such that salience and absoluteness are systematically teased apart. We do this by varying the salience of size between experiments. Throughout the experimental series, we also explore the tendency toward consistent behavior, examining to what extent speakers alternate between including and not including an attribute, and investigating the effect of including necessary attributes on the production of size overspecification in particular.

Experiment 1 is a baseline study in which we investigate the rates of color, pattern, and size overspecification. As discussed in the previous section, color, which has been argued to be "special" with respect to overspecification, is similar to pattern in being salient and absolute (see **Table 1**). Size, on the other hand, differs from color and pattern in being relative instead of absolute. Further, in Experiment 1, the contrast between big and small items is low and size is hence low in salience. As such, size is different from both color and pattern, in being relative and less salient. If speakers tend to include color because it is salient and absolute, they are expected to include other attributes that are salient and absolute as well. We therefore hypothesize that in comparison to size overspecification, speakers will not only produce more color overspecification, which would be in line with what has been found before (Pechmann, 1989; Belke and Meyer, 2002; Gatt et al., 2013), but also more pattern overspecification.

In Experiment 2, we explore the possibility that in Experiment 1, where a within-participants design is used, the expected tendency toward consistency may lead to an effect of the tendency to include or not include one attribute on the rate of overspecification of another attribute. For example, pattern might be treated like color because the two attributes share characteristics with each other but not with size. Another possibility is that not including size in their utterances will lead some speakers to stop producing color and pattern overspecification as well. In Experiment 2, we investigate the occurrence of such effects in Experiment 1, by varying the three attributes between instead of within participants. If the rates of overspecification tend to affect one another, the pattern of

TABLE 1 | Salience and absoluteness of the three attributes.

	Salience	Absolute
Color	High	Yes
Pattern	High	Yes
Size	Experiments 1 and 2: Low Experiments 3 and 4: High	No

[3] To be precise, the discriminatory power of a referent's attribute is computed by dividing the number of competitors (the objects in the visual context other than the referent) that do not share the attribute with the referent by the total number of competitors.

results is expected to change compared to the pattern found in Experiment 1.

In Experiment 3, we delve into the question of how salience and absoluteness contribute to the tendency to include attributes, teasing these two features apart. We make size more salient by increasing the contrast between big and small items. We hypothesize that the rate of size overspecification increases correspondingly, which would indicate that salience is a factor in selecting attributes and producing overspecification. Furthermore, we expect absoluteness to have an effect, too, leading to higher rates of overspecification of the two absolute attributes (color and pattern) than the relative attribute (size).

Experiment 4, finally, investigates whether overspecification of the three attributes is triggered by including non-critical trials which, unlike the critical trials, require color, pattern, or size to be included. The experiment is thus conducted to assess whether the production of overspecification of color, pattern, and even size, can increase due to a tendency toward consistency.

3. EXPERIMENT 1

In Experiment 1, we vary color, pattern, and size in a within-participants design and compare the rates of overspecification for the three attributes. As color and pattern are salient and absolute while size is less salient and relative, we hypothesize that the rates of color and pattern overspecification will be higher than the rate of size overspecification. We also explore the tendency toward consistency by examining the individual proportions of alternations between overspecification and minimal specification in each condition.

3.1. Method

3.1.1. Participants

We tested 18 native speakers of Dutch (14 females, 4 males, mean age 23 years, range 18–27 years) at Radboud University, Nijmegen, the Netherlands. All were volunteers and they received a small fee for their participation. All of them reported not to be colorblind.

3.1.2. Materials

We used six line drawings of clothes as stimulus materials, which were collected on Google Image. All garments would normally be named by a one-syllabic noun in Dutch. The six pictures were manipulated in order to create variation on the three attributes. Relative size is expressed in Dutch by equivalents of “big” and “small,” which makes it basically a binary attribute. We therefore selected two values of each of the two other attributes, too. The pattern values were striped and spotted, the color values were blue and green, and the size values were big and small, as shown in **Figures 1–3**. We thus created six variants of each picture. The patterns were clear gray stripes or spots against a white background and the colors were bright, saturated colors. The ratio between the heights of the big and small pictures was 3:2. The experiment was programmed with Presentation software.

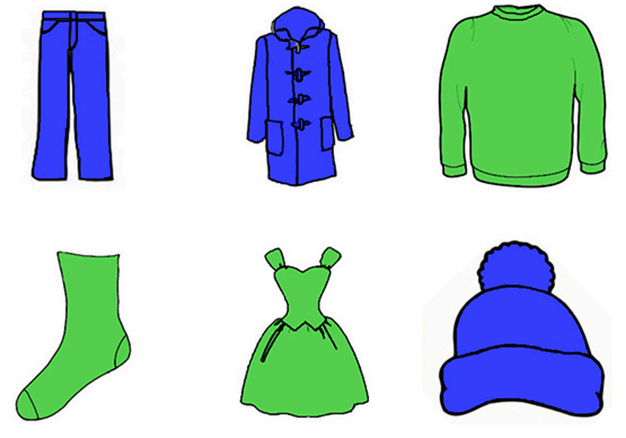


FIGURE 1 | An array in the Color condition in Experiment 1.

We also had filler pictures, which were taken from the Tarrlab Stimulus Repository[4]. There were three types of filler pictures: common objects, like bikes and envelopes (Rossion and Pourtois, 2004), Greebles (Gauthier and Tarr, 1997), and human faces. Greebles are complex and visually similar, which makes them difficult to describe uniquely. So as not to stimulate participants to pay special attention to color, filler pictures were presented in desaturated, inconspicuous colors (common objects) or in gray tones (Greebles).

3.1.3. Design

In critical trials, an array was presented with pictures of six different garments. They were arranged in a 2 (row) × 3 (column) grid. We had three conditions: Color, Pattern, and Size. The objects within an array always varied on exactly one attribute: color, pattern, or size, respectively. In each array, half of the objects had one value (e.g., striped) and the other half had the other value (e.g., spotted). The target object thus shared its value with two other objects. Including a color, pattern or size modifier always resulted in overspecification. Examples of arrays are shown in **Figures 1–3**.

Attribute was manipulated within participants: each participant received trials from all three conditions. Each of the six objects once acted as target in each of the six possible values, yielding 36 critical trials. All participants saw all critical trials. They also saw 36 trials of each of the three filler types, yielding a total of 144 trials. Eight additional trials were included for practice.

Fillers were included for two reasons: first, to prevent participants from sticking to one syntactic and semantic structure throughout the whole experiment, and second, to hide the purpose of the experiment. There were three types of filler trials. Fillers of the first type consisted of arrays with four pictures of common objects, which were included to elicit unmodified referring expressions, that is, expressions without

[4] Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/. In some cases, colors were adjusted or images were mirrored.

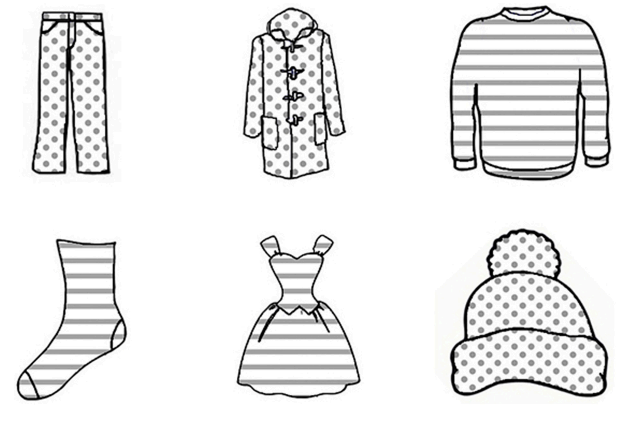


FIGURE 2 | An array in the Pattern condition in Experiment 1.

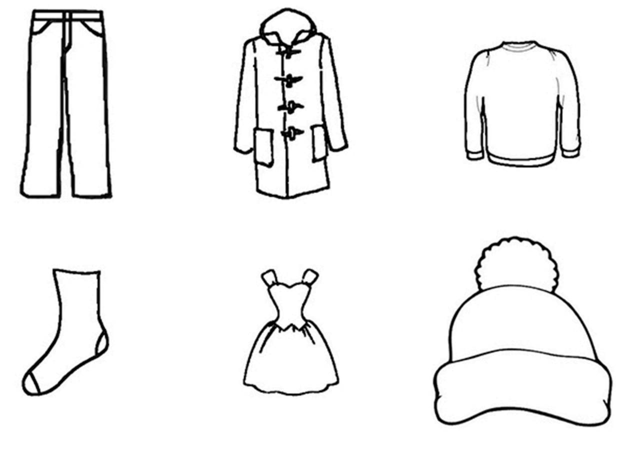


FIGURE 3 | An array in the Size condition in Experiment 1.

any adjectives or prepositional phrases. We did not expect modification to occur because basic-level terms were always sufficient and pictures did not have striking or unexpected features. Fillers of the second type were arrays with four pictures of Greebles, which were included to make participants aware that simply naming objects was not always sufficient. Fillers of the third type were arrays with two human faces, which were either of the same gender or of different genders. They were included to elicit variation in the presence of modifiers within a category: modification was necessary when the two people were of the same gender, but unnecessary when they differed in gender.

The order of the trials was pseudorandomised, with the restriction that a trial was always followed by at least two trials in which the target was of a different type of garment. For example, when the target was a sock, the target in the next two trials was never a sock. We did this in order to prevent participants from producing an adjective for the sake of contrast between the referent and the previous referent, which speakers have been shown to do in reference production experiments (see Levelt, 1989, p. 132; Pechmann, 1989, for discussion of this type of factors in reference production). Each participant saw the trials in a unique order.

3.1.4. Procedure

Participants were tested individually in a quiet booth. Their task was to instruct an imaginary addressee to click on one of the pictures, by completing the Dutch equivalent of the sentence "Click on ...." A cross preceding the array indicated the position of the target on the screen. Participants were asked to formulate their instruction in such a way that an addressee would be able to click on the right picture, even if the pictures would be arranged differently on the screen for the addressee than for the participant. This particular instruction was given to prevent them from referring to the location of the pictures on the screen. It took participants about 20 min to complete the task.

3.2. Results

Each participant performed 36 critical trials. In two trials, no response was given. The critical trials thus elicited 646 responses. Seventeen responses (2.6%) were removed, because the referent was not the target item, or because the speaker corrected themselves during the articulation of the utterance. The remaining 629 expressions were annotated as overspecified when a color modifier was included in the Color condition, when a pattern modifier was included in the Pattern condition, and when a size modifier was included in the Size condition[5].

Experiment 1 was conducted to answer the question how likely speakers are to produce overspecification of color, pattern, and size, respectively. We expected that overspecification would be produced more often in the Color and the Pattern conditions than in the Size condition. Indeed, **Figure 4** shows that overspecification was produced often in the Color condition (proportion of overspecification: $M = 0.55$, $SD = 0.50$) and in the Pattern condition ($M = 0.42$, $SD = 0.49$), but almost never in the Size condition ($M = 0.01$, $SD = 0.10$).

In this experiment and all the following, Shapiro-Wilk tests indicated that the data were not normally distibuted ($p < 0.001$ in all conditions in all experiments). Hence, we ranked the data and used non-parametric statistics for the analyses. We report mean ranks, denoted by *MR*.

A Friedman's ANOVA indicated a highly significant main effect of Attribute on overspecification, $\chi^2(2) = 24.24, p < 0.001$. In line with our hypothesis, stepwise stepdown comparisons indicated a significant difference between the Pattern ($MR = 2.17$) and Size ($MR = 1.19$) conditions, $p = 0.005$, while the difference between the Pattern and the Color ($MR = 2.64$) conditions was not significant, $p > 0.1$.

To explore the tendency toward consistent behavior, we counted the number of times that participants included an attribute in a trial but did not include it in the next trial of the same condition, or vice versa. For each participant, we divided

[5] This means we did not take into account *all* occurrences of overspecification. Color was sometimes included in the Pattern condition ($n = 9$) or in the Size condition ($n = 2$), but we did not count these cases as color overspecification. Doing so would not have yielded a fair comparison between the attributes because only pictures in the Pattern condition had patterns, while all pictures had a color. Moreover, patterns in line drawings are only there by the grace of color contrast.

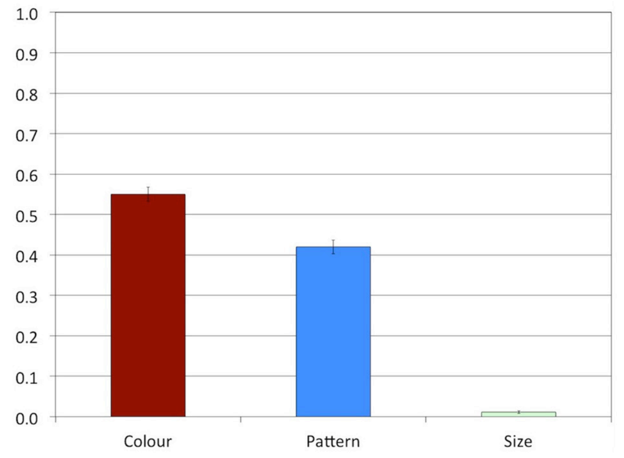


FIGURE 4 | Experiment 1: Proportions of overspecified referring expressions. The error bars represent standard errors.

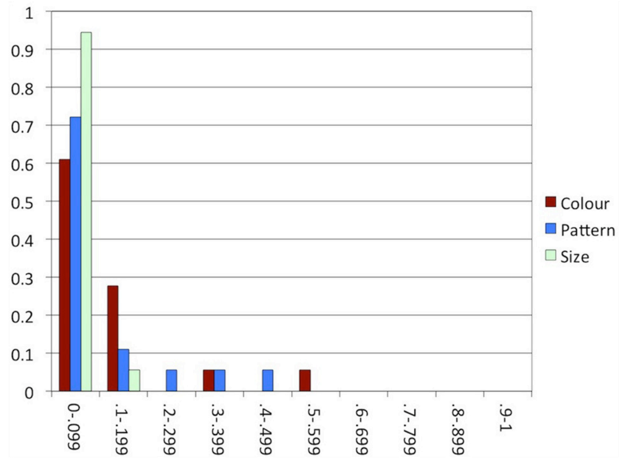


FIGURE 5 | Experiment 1: The proportion of participants (y-axis) in each range of proportions of alternations in each condition (x-axis).

this number by the number of trials of the condition −1 (the number of opportunities to alternate). **Figure 5** shows the degree of consistency in each condition, indicating that participants tended to behave highly consistently, the majority alternating in less than 10% of the trials within each condition.

3.3. Discussion

Experiment 1 indicates that, in line with our expectations, speakers produced substantial rates of color and pattern overspecification, but hardly any size overspecification. Although the rate of color overspecification was numerically higher than the rate of pattern overspecification, this difference was not significant. It seems, then, that color and pattern overspecification are both likely to occur, both attributes being salient and absolute. In line with the literature, we found that speakers were highly consistent within conditions, most of them either producing or not producing overspecification in the majority of the trials.

As was pointed out before, the tendencies to include or not include one attribute may have affected the rate of overspecification of another attribute, due to a tendency toward consistency. It is possible, for example, that a tendency to include color may have triggered the production pattern overspecification, since the two attributes share characteristics with each other but not with size. Another possibility is that the tendency not to include size has resulted in a decrease in overspecification overall.

In Experiment 2, we vary the three attributes between participants, thereby excluding the possibility that the rate of overspecification in one condition affects the rate in another. A change in the pattern of results would therefore indicate that such between-attributes effects took place in Experiment 1, probably due to the tendency toward consistency. A stable pattern, in contrast, would show that the rates of overspecification of the three attributes did not affect one another.

4. EXPERIMENT 2

In Experiment 2, we vary color, pattern, and size in a between-participants design, in order to find out whether the rates of overspecification in Experiment 1 affected one another, due to a tendency toward consistent behavior. A change in the pattern of results would indicate that such effects occurred, whereas a similar pattern would show that they were absent. Again, we expect a high degree of consistency within speakers.

4.1. Method

4.1.1. Participants

We tested 54 participants (43 females, 11 males, mean age 22 years, range 18–31 years) similar to those in Experiment 1[6]. None had participated in the previous experiment.

4.1.2. Materials, Design, and Procedure

Materials were the same as in Experiment 1. Attribute was now manipulated between participants. Participants were randomly assigned to either of the three conditions: Color, Pattern, or Size, with 18 participants per group. In each condition, there were twelve different critical pictures in each condition (6 pictures * 2 values of the attribute in that condition). Each picture was presented twice in each experimental session, yielding 24 critical trials in each condition. Participants also received 24 trials of each of the three filler types, yielding a total of 96 trials. Four additional trials were included for practice. Otherwise design and procedure were the same as in Experiment 1.

4.2. Results

All participants performed 24 critical trials. Once, no response was given. The critical trials thus elicited 1295 responses. We excluded 28 responses (2.2%) from the analysis as in

[6] Data from eight additional participants were collected but not analyzed because they were instructed incorrectly ($n = 4$), because they received the wrong practice trials ($n = 1$), or because they failed to produce definite descriptions ($n = 3$).

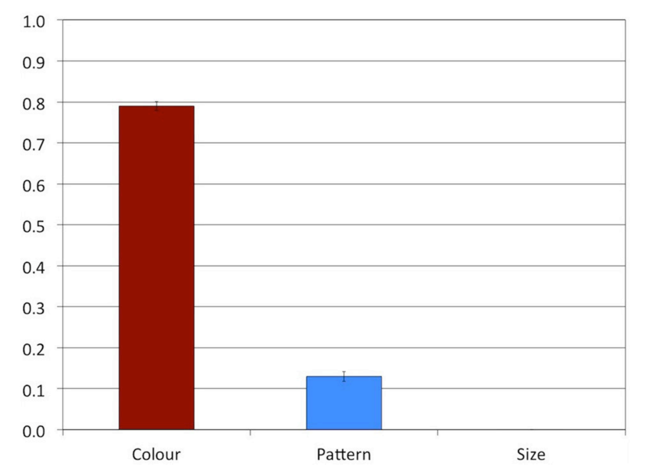


FIGURE 6 | Experiment 2: Proportions of overspecified referring expressions. The error bars represent standard errors.

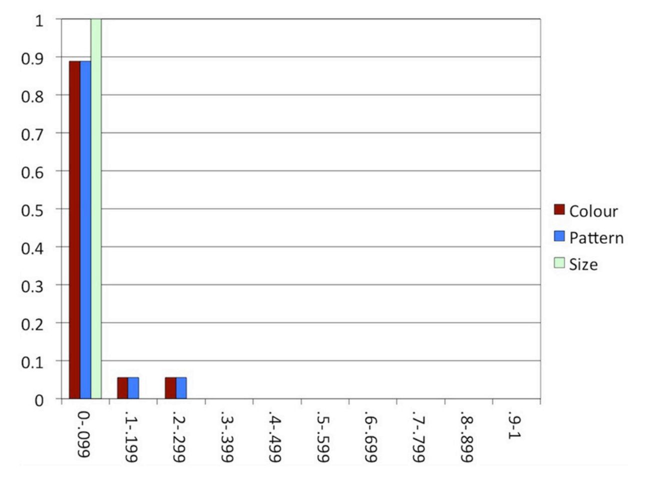


FIGURE 7 | Experiment 2: The proportion of participants (y-axis) in each range of proportions of alternations in each condition (x-axis).

Experiment 1. The remaining 1267 expressions were annotated as in Experiment 1[7].

A comparison of **Figures 4**, **6** suggests that the patterns of results found in Experiments 1 and 2 were different, indicating that varying the three attributes within participants affected the proportions of overspecification in Experiment 1. A Kruskall-Wallis test indicated a main effect of Attribute in Experiment 2, $H(2) = 35.98$, $p < 0.001$. Stepwise stepdown comparisons revealed that the proportion of overspecification was significantly higher in the Color condition ($M = 0.79$, $SD = 0.41$, $MR = 42.94$) than in the Pattern condition ($M = 0.13$, $SD = 0.34$, $MR = 22.06$), $p < 0.001$. Although overspecification in the Size condition was at floor, it was still significantly lower ($MR = 17.50$) than in the Pattern condition, $p = 0.037$.

A Mann-Whitney test showed that the rate of pattern overspecification was significantly lower in Experiment 2 ($MR = 14.33$) than in Experiment 1 ($MR = 22.67$), $U = 87.00$, $z = 2.61$, $p = 0.017$, which indicates that the rate of pattern overspecification in Experiment 1 was affected by the tendencies to include or not include the other attributes. The rate of color overspecification was numerically higher in Experiment 2 ($MR = 21.72$) than in Experiment 1 ($MR = 15.28$), but this difference was only marginally significant, $U = 220.00$, $z = 1.91$, $p = 0.07$.

As in Experiment 1, most participants alternated between producing and not producing overspecification within conditions in less than 10% of the trials, as indicated in **Figure 7**. That is, consistency was high again, which is in line with our expectation.

4.3. Discussion

The patterns of results found in Experiments 1 and 2 were clearly different, indicating that the rates of overspecification in Experiment 1 affected one another. In contrast to what was found in Experiment 1, where the rates of color and pattern overspecification were statistically indistinguishable, there was a large and highly significant difference between the Pattern and the Color conditions in Experiment 2. Although the rate of overspecification was significantly higher in the Pattern than in the Size condition in both experiments, the rate of pattern overspecification was closer to the rate of color than to the rate of size overspecification in Experiment 1, while it was the other way around in Experiment 2. A significant difference between the two Pattern conditions in Experiments 1 and 2 suggests that the production of color overspecification in Experiment 1 triggered the production of pattern overspecification. We found no evidence, on the other hand, that color overspecification *decreased* due to a tendency to *not* produce size overspecification: although the rate of color overspecification was numerically higher in Experiment 2 than in Experiment 1, this difference did not reach significance.

Experiment 2 indicates that the tendency to include color is stronger than the tendency to include pattern. Since both attributes are absolute, a possible explanation is that pattern is less salient than color. On the other hand, while the tendency to produce color overspecification may have triggered some participants to produce pattern overspecification, it did not trigger them to produce size overspecification. This may be because size is still less salient than pattern, but it may also be due to the fact that size is a relative attribute while both color and pattern are absolute.

In Experiment 3, we vary the three attributes within participants again, and we increase the contrast between big and small items, making size more salient. This enables us to investigate the respective effects of salience and absoluteness on the tendency to include attributes. In line with van Gompel et al. (2014), we might expect the rate of size overspecification to increase, indicating that salience is a factor in the tendency to include attributes and to produce overspecification. Furthermore, we expect an effect of absoluteness, resulting in a difference between color and pattern on the one hand, and size on the other hand, as in Experiment 1.

[7] Participants never mentioned color in the Pattern or Size conditions, as happened sometimes in Experiment 1.

5. EXPERIMENT 3

In Experiment 3, we assess how salience and absoluteness contribute to the tendency to select attributes in referring expressions. As in Experiment 1, we vary color, pattern, and size within participants, but now increasing the salience of size, in order to find out whether this results in an increase in size overspecification compared to Experiment 1, which would indicate an effect of salience on overspecification. We also expect that there will remain a difference between the two absolute attributes (color and pattern) and size. Finally, we expect the degree of consistency within speakers again to be high.

5.1. Method

5.1.1. Participants

We tested 18 participants (13 females, 5 males, mean age 21 years, range 18–29 years) similar to those in the previous experiments. None had participated in either of the previous experiments.

5.1.2. Materials and Design

In the Size condition, the ratio between big and small pictures was 3:1 instead of 3:2. An example of an array in the Size condition is shown in **Figure 8**. Otherwise, materials, design, and procedure were as in Experiment 1.

5.2. Results

All participants performed 36 critical trials each. Once, no response was given. The critical trials thus elicited 647 responses. Seven responses (1.1%) were removed from the analysis as in Experiment 1. The remaining 640 responses were annotated as in the previous experiments.

We conducted Experiment 3 to assess how salience and absoluteness contribute to the tendency to select attributes. Our first hypothesis was that an increase in salience of size would result in an increase in the rate of size overspecification from Experiment 1 to 3, indicating that salience contributes to this tendency. We also expected absoluteness to contribute, our second hypothesis being that there would still be a difference between color and pattern on the one hand, and size on the other hand (like in Experiments 1 and 2).

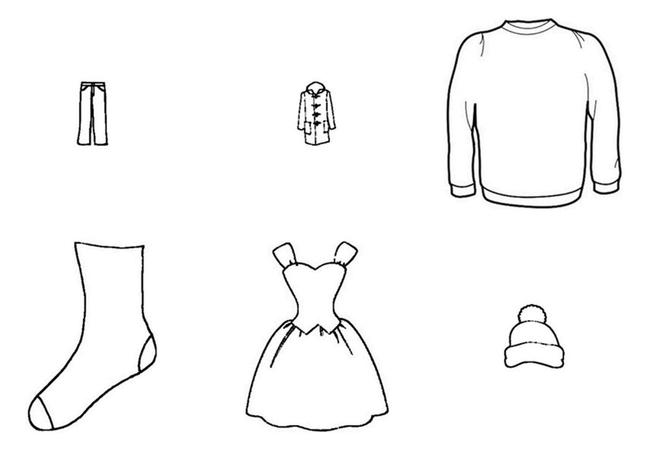


FIGURE 8 | An array in the Size condition in Experiment 3.

The proportions of overspecified referring expressions in each condition in Experiment 3 are shown in **Figure 9**. A Mann-Whitney test indicated that although the proportion of size overspecification was numerically higher in Experiment 3 ($M = 0.11$, $SD = 0.31$, $MR = 20.17$) than in Experiment 1 ($M = 0.01$, $SD = 0.10$, $MR = 16.83$), this difference was not significant, $U = 129.00$, $z = 1.38$, $p > 0.1$. Thus, our first hypothesis was not confirmed by the data.

In line with our second hypothesis, **Figure 9** suggests that the patterns of Experiments 1 and 3 were globally similar, with overspecification being produced more often in the Color and the Pattern conditions than in the Size condition. Two additional Mann-Whitney tests confirmed that there was no significant difference between Experiments 1 and 3 for Color ($MR = 20.72$ vs. $MR = 16.28$, $U = 122.00$, $z = -1.28$, $p > 0.1$), and for Pattern ($MR = 20.08$ vs. $MR = 16.92$, $U = 133.50$, $z = -0.94$, $p > 0.1$).

A Friedman's ANOVA indicated that there was a significant main effect of Attribute, $\chi^2(2) = 19.58$, $p < 0.001$. Stepwise stepdown comparisons showed that the difference between the Color ($M = 0.37$, $SD = 0.48$, $MR = 2.56$) and Pattern ($M = 0.29$, $SD = 0.45$, $MR = 2.03$) conditions was not significant, $p > 0.10$, as in Experiment 1, and that the difference between Pattern and Size ($MR = 1.42$) was marginally significant, $p = 0.059$.

Earlier, we found a significant difference between the two Pattern conditions in Experiments 1 and 2, while the difference between the two Color conditions was only marginally significant (see Section 4.2). We thus found evidence that in Experiment 1, the rate of pattern overspecification was affected by tendencies to include or not include other attributes, but no evidence for analogous effects on the rate of color overspecification. However, a Mann-Whitney test indicates that the proportion of color overspecification was significantly lower in Experiment 3 ($M = 0.37$, $MR = 13.86$) than in Experiment 2 ($M = 0.79$, $MR = 23.14$), $U = 78.50$, $z = -2.71$, $p = 0.007$, indicating that the rate of color

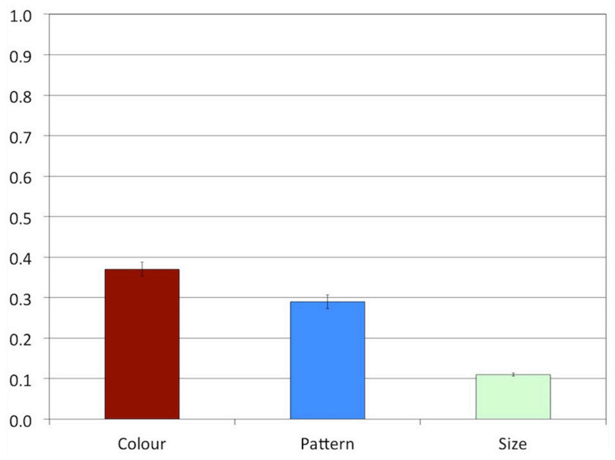


FIGURE 9 | Experiment 3: Proportions of overspecified referring expressions. The error bars represent standard errors.

overspecification, too, is affected by the way other attributes are treated.

As in the previous studies, most participants alternated between producing and avoiding overspecification within conditions in less than 10% of the trials, as indicated in **Figure 10**. That is, consistency was high again, which is in line with our expectation.

5.3. Discussion

Experiment 3 was conducted to assess how salience and absoluteness contribute to the tendency to produce overspecification of color, pattern, and size. We hypothesized that due to an increase in salience, the rate of size overspecification might increase, but that due to a difference in absoluteness, the rates of color and pattern overspecification would remain higher than the rate of size overspecification.

Our first expectation was not confirmed: there was no significant difference between the rates of size overspecification in Experiments 1 and 3. At first sight, this result does not seem to be in line with the findings of van Gompel et al. (2014), who did find a positive effect of increasing salience of size on size overspecification. However, as discussed in Section 2.2.1, there is a crucial difference between their experiments and ours: in their study, all items were of the same type but different sizes and colors, requiring either size or color for disambiguation between the target and the other objects, while in our study, all items were of different types and therefore it was never necessary to add a modifier to the noun. Thus, including size resulted in overspecification in our study, while in theirs, only including both color and size did. Even if both color and size were included in their study, however, size might still not be experienced as irrelevant by an addressee, because it did distinguish between objects of the same type. An eyetracking study conducted by Sedivy et al. (1999), which was touched upon briefly in Section 2.1, indicated that addressees expect speakers to use size adjectives only if the referent has a bigger or smaller counterpart in the context, whereas they do not have analogous expectations about the use of color adjectives. In this study, participants were shown arrays with, for example, a big and a small glass, a big pitcher, and a small key. Eye gaze patterns suggested that upon hearing "big," participants inferred that the referent was the big glass rather than the big pitcher, whereas in a situation with a pink and a yellow comb, a yellow bowl, and a knife, they did not infer from hearing "yellow" that the referent was the yellow comb rather than the yellow bowl. These findings suggest that size adjectives are expected only if there is a relevant size contrast in the context, that is, if the referent is bigger or smaller than another object of the same type. There was such a relevant size contrast in the experiment conducted by van Gompel et al., where all objects in the array were of the same type, but not in our experiments, where all objects were of different types. Size overspecification would therefore violate an addressee's expectation, and possibly even lead to confusion, when produced in the visual contexts we used in our experiment, but not in the contexts used in van Gompel et al.'s study. This may urge speakers to avoid size overspecification when there is no relevant size contrast in the context, probably due to the fact that size is a relative attribute.

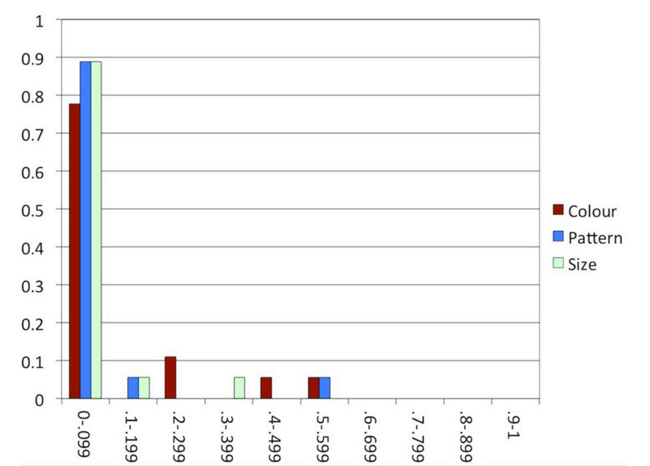


FIGURE 10 | Experiment 3: The proportion of participants (y-axis) in each range of proportions of alternations in each condition (x-axis).

Alternatively, it is possible that the difference between van Gompel et al.'s findings and ours is due to the fact that the size contrast in their study was 5:1 whereas it was 3:1 in our study. As **Figure 8** shows, however, the size contrast in our study was quite striking, which led one of the participants in a pilot study to ask for "the *very* small dress" ("de *hele* kleine jurk") the first time when she came across a trial in which a small object was the target. We therefore think it unlikely that participants in our study did not include size because it was not sufficiently salient.

The absence of a significant effect of salience on size overspecification and the difference between our results and those found by van Gompel et al. suggest that absoluteness is an important factor in attribute selection: even if size is made salient, size overspecification is produced infrequently. This suggestion is in line with our expectation that due to the difference in the absoluteness dimension, the rate of size overspecification would remain lower than the rates of color and pattern overspecification. Although the difference between pattern and size was only marginally significant, we did find that the pattern of results in Experiment 3 was globally similar to the one in Experiment 1, where this difference was highly significant. None of the three conditions in Experiment 3 was significantly different from the corresponding conditions in Experiment 1. Besides, in both experiments, proportions of overspecification in the Color and Pattern conditions were statistically indistinguishable, and they were numerically closer to each other than either of them was to the Size condition. All in all, this suggests that absoluteness indeed contributes to the tendency to include certain attributes but not others.

If the low frequency of size overspecification in Experiment 3 is indeed due to the fact that the contrast on this relative attribute was irrelevant, this may also explain why the rate of color overspecification in Experiment 3 was so much lower than in Experiment 2. We know from the previous experiments that speakers strongly tend to behave consistently, treating similar attributes in a similar way. In Experiment 1, this resulted in the majority of participants including both color and pattern but not size, which was different from the other two in being relative and low in salience. The high salience of size in Experiment 3, however, may have led participants to treat all three attributes similarly, since all of them were salient, either including them all or including none of them. Since including them all would lead to the unnecessary and irrelevant mention of a relative attribute, the majority of the participants may have been triggered to produce no overspecification at all.

It might be noted that, as in the previous experiments, our manipulation of the size of the pictures was independent of the proportions among the objects that the pictures represent: for example, a dress is normally much larger than a sock. Because people are so experienced in interpreting pictures and their sizes, which are not always proportional to real life sizes, we assume that our participants will have had no problem interpreting the size of the pictures in the arrays. Letting go of real life proportions was inevitable in the light of our purpose, namely, to compare the rates of overspecification of size with the other two attributes. In many other studies (such as van Gompel et al.'s), size differences are indicated by representing several objects of the same type in different sizes (for instance, a small candle and several larger candles). As discussed in Section 2.2.1, this is suitable when the *competition* between size and other attributes is investigated: how likely are speakers to include size when including either size or color is sufficient? In that situation, overspecification only arises when both size and color are included. In the present study, however, we are interested in a comparison between overspecification of different attributes, including size. To investigate this, it is necessary that the target object is unique in a display and that it differs in size from different objects. As it is hard, if not impossible, to indicate in a realistic way that a sock is small *for a sock* by exploiting the proportion between the sock and a dress, especially if the ratio between big and small pictures is fixed, we decided to abstract from the natural sizes of the objects represented. The fact that size overspecification was often produced in Experiment 4 (see Section 6.2), in which the same displays were used, indicates that it is unlikely that participants were confused by the "unnatural" size differences between the pictures.

Experiment 3 shows that size overspecification is produced infrequently if there is no relevant size contrast in the visual context, even if size is made highly salient. In Experiment 4, we investigate whether there are nevertheless circumstances that *do* trigger size overspecification, even if there is no relevant size contrast. As speakers show a tendency toward consistency, triggering the mention of the three attributes is likely to result in an increase in the rates of overspecification of all attributes, including size.

6. EXPERIMENT 4

In Experiment 4, we investigate circumstances that may trigger size overspecification, by introducing non-critical trials which require speakers to include color, pattern, or size in order to yield a unique description. Since participants in previous studies were found to show a strong tendency toward consistency, we expect the non-critical trials to trigger mentioning the three attributes, yielding an increase in color and pattern in comparison with Experiment 3, and also, for the first time, the occurrence of size overspecification, even though there is no relevant size contrast present in the visual context.

6.1. Method

6.1.1. Participants

We tested 20 participants (16 females, 4 males, mean age 22 years and 10 months, range 18–28 years) similar to those in Experiment 1[8]. None had participated in any of the previous experiments.

6.1.2. Materials, Design, and Procedure

The critical pictures used in Experiment 3 were now used both as critical and non-critical pictures. The pictures that were used as fillers in the previous experiments were not used here. Otherwise, materials and procedure were as in the previous experiments.

As in Experiment 3, attribute was manipulated within participants. Non-critical trials were now included to trigger the use of modifiers. They were identical to critical trials, except that one of the garments shared the target's type (but not its value). For example, when the target was a big sock, then there was also a small sock in the array. In this context, omitting a size modifier ("Click on the sock") would result in underspecification, which we know from a variety of studies to be rarely produced (e.g., Engelhardt et al., 2006; Arts et al., 2011b; Koolen et al., 2011; Davies and Katsos, 2013). Additionally, in half of the trials discriminatory power was increased to make the target value more salient and hence increasing the probability that speakers would include size modifiers even in the critical trials. In half of the trials, as in the previous experiments (LowDist), the target shared its value with two of its distractors (see **Figures 1**–**3**), whereas in the other half (HighDist), it did not share its value with any of them, increasing this value's salience. For example, if the target in the HighDist condition was blue, the five other pictures were green.

All 36 variants of each picture acted as the target of a critical trial twice: they acted as target once in the LowDist condition and once in the HighDist condition. They also acted as the target of a non-critical trial twice, yielding a total of 144 trials. Six additional trials were included for practice.

6.2. Results

All participants performed 72 critical trials each. The critical trials elicited 1440 responses, 45 of which (3.1%) were excluded from the analysis as in Experiment 1. The remaining 1395 were annotated as in Experiment 1.

Experiment 4 was conducted to answer the question whether even size overspecification is triggered by mentioning color, pattern, and size. Additionally, in half of the critical trials (HighDist condition), we increased the salience of the target's value by making it unique in the array. The proportions of overspecified referring expressions in each condition are shown

[8] Data from two additional participants were collected but not analyzed because they did not follow the instructions ($n = 1$) or because their age exceeded the upper age bound of 35 ($n = 1$).

in **Figure 11**. In all conditions, including the Size condition, the proportion of overspecified referring expressions was now strikingly high, namely between 0.7 and 0.8. A comparison with the results of Experiment 3, presented in **Figure 9**, indicates an increase in the rate of color and pattern overspecification, and, crucially, also of size overspecification.

A Wilcoxon Signed Rank test was conducted first, to find out whether discriminatory power had an effect on overspecification. This turned out not to be the case, $z = 1.28$, $p = 0.20$, $r = 0.29$. Hence, the HighDist and the LowDist conditions were collapsed in all subsequent analyses.

Indeed, a Mann-Whitney test confirmed that the difference between Experiments 3 and 4 was highly significant for the Size conditions ($MR = 11.11$ vs. $MR = 27.05$, $U = 331.00$, $z = 4.54$, $p < 0.001$), and also for the Color ($MR = 13.50$ vs. $MR = 24.90$, $U = 288.00$, $z = 3.26$, $p = 0.001$) and the Pattern conditions ($MR = 14.14$ vs. $MR = 24.32$, $U = 276.50$, $z = 2.97$, $p = 0.004$).

Finally, a Friedman's ANOVA indicated that there was a significant main effect of Attribute in Experiment 4, $\chi^2(2) = 11.81$, $p = 0.003$. Pairwise comparisons indicated that the differences between Color ($MR = 2.40$) and Pattern ($MR = 2.08$) and between Pattern and Size ($MR = 1.52$) were not significant, $p > 0.08$ for both comparisons, while the difference between Color and Size was significant, $p = 0.006$.

As indicated in **Figure 12**, consistency was high, as in all previous experiments. In line with our expectation, the majority of the participants produced or avoided overspecification most of the time in all conditions.

6.3. Discussion

Experiment 4 shows that the strong tendency not to produce size overspecification that we found in our previous experiments can disappear almost entirely when mentioning color, pattern, and size is triggered. Although even in this experiment, more overspecification was produced in the Color than in the Size condition, the proportion of size overspecification strongly increased due to the non-critical trials, which required size modifiers, and it was very close to the proportions of overspecification in the Color and Pattern conditions, which were also significantly higher than the proportions of their counterpart conditions in Experiment 3.

To conclude, Experiment 4 provides evidence that overspecification, even of size, can be triggered under certain circumstances, due to a general tendency to behave consistently. Speakers thus do not necessarily avoid overspecification of a relative attribute, even if there is no relevant contrast on this attribute in the visual context.

7. GENERAL DISCUSSION

In this paper, we investigated the tendencies to produce color, pattern, and size overspecification. We compared rates of overspecification of the three attributes, focusing on the role of salience, absoluteness, and consistency. Since color and pattern are salient and absolute whereas size is relative and often less salient, we hypothesized that speakers would produce more color and pattern overspecification than size overspecification. Experiment 1, which had a within-participants design, confirmed this expectation: speakers produced substantial rates of color and pattern overspecification, which were very similar to each other, but almost no size overspecification.

Experiment 2 indicated, however, that in Experiment 1, pattern was treated similarly to color because the rates of overspecification affected one another: when varying the attributes between participants, the proportion of pattern overspecification was low, while the proportion of color overspecification was high. The tendency to select pattern is thus less strong than the tendency to select color. As both are absolute attributes, a possible explanation for this finding is that pattern is less salient than color. We concluded that in Experiment 1, the tendency to produce color overspecification probably stimulated the production of pattern overspecification, which is likely to be due to the fact that the two attributes are absolute and more

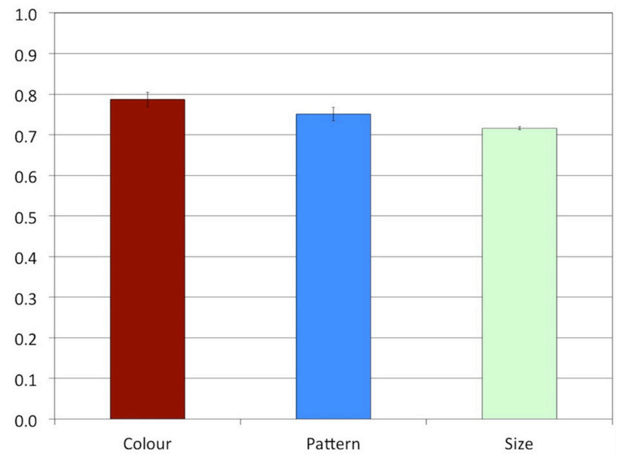


FIGURE 11 | Experiment 4: Proportions of overspecified referring expressions. The error bars represent standard errors.

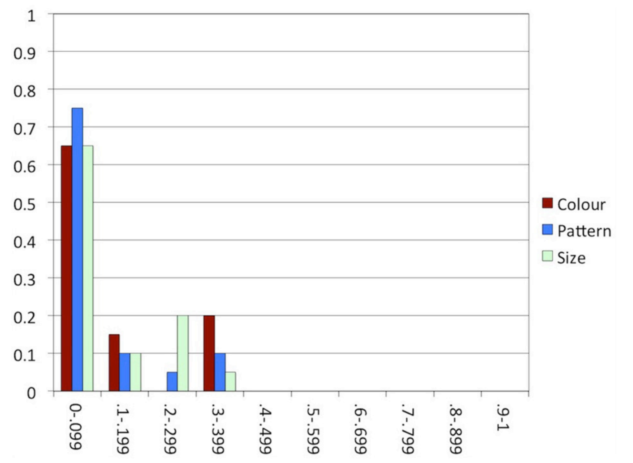


FIGURE 12 | Experiment 4: The proportion of participants (y-axis) in each range of proportions of alternations in each condition (x-axis).

salient than size. A comparison between Experiments 2 and 3, in which the three attributes were manipulated within participants again, indicated that the rates of overspecification of the three attributes can also affect one another in a different way: the rate of color overspecification was significantly lower in Experiment 3 than in Experiment 2. A plausible explanation is that the tendency *not* to include size triggered some participants to not include color either. In sum, Experiment 2 shows that the rates of overspecification of different attributes can affect one another due to a tendency toward consistency.

Experiment 3 was conducted to assess how salience and absoluteness contribute to the tendencies to select attributes. As in Experiment 1, attribute was manipulated within participants, but size was now made more salient by increasing size contrast. This manipulation did not result in a significant increase in size overspecification, however, and the patterns found in Experiments 1 and 3 were globally similar. In contrast to our findings, van Gompel et al. (2014) found that an increase in size contrast made speakers stop preferring color over size. Importantly, the size contrast in their study was relevant: when the referent was a small candle, there were also large candles in the array. In our study, in contrast, the referent was always unique, and the size contrast was therefore not relevant. Thus, an increase in salience can trigger selection of size, as van Gompel et al. show, but our study shows that salience is not *enough* to trigger size selection. The fact that a relevant contrast in the context seems to be crucial for including size suggests that size overspecification is infrequent because size is a relative attribute, indicating that absoluteness is a factor in attribute selection. This was supported by the fact that the pattern of results found in Experiment 3 was globally similar to the one in Experiment 1, where color and pattern were treated similarly, and differently from size, even though the difference between pattern and size was only marginally significant in Experiment 3.

In Experiment 4, finally, we found that even size overspecification can be triggered by mentioning color, pattern, and size, even though there was no relevant size contrast present in the critical trials. This finding is in line with Goudbeek and Krahmer (2012), who found that the selection of dispreferred attributes can be primed. It shows that the strong tendency toward consistency that was also found in the other three experiments can even lead to overspecification of attributes which otherwise do not tend to be included redundantly.

In many earlier studies investigating consistency in reference production, speakers appeared to have good reason to switch to a different construction: in Brennan and Clark (1996) and Van Der Wege (2009), the modified or otherwise highly specific terms that had been used before in the discourse would normally be dispreferred in the new context, and the attributes primed in Goudbeek and Krahmer (2012) are known to be normally dispreferred, too. The arrays used in critical trials in our experiments, in contrast, were highly similar, providing little reason for alternating between overspecification and minimal specification within conditions. This is especially clear in Experiment 2, where for each individual participant, objects in all arrays varied in the same attribute. Indeed, comparing **Figures 5**, **7**, **10**, **12** suggests that consistency was highest in Experiment 2. In the other experiments, where attribute was manipulated within participants, the alternation of the three attributes may have enhanced alternating between including and not including attributes within conditions.

As was stated in the Introduction, we are neutral as to what mechanisms underpin the tendency toward consistency in reference production in our experiments, and our study was not meant to settle the debate on those mechanisms. Still, it is worth pointing out that we think it most likely that the consistent behavior we found was due to priming. Although it is not impossible that our participants sought to establish conceptual pacts with their imaginary hearer, experimental studies suggest that effects of common ground considerations are so subtle that they can only be detected when the experimental set-up is sufficiently natural. For example, Brown and Dell (1978) seemed to show that interlocutors do not routinely take into account the common ground when telling stories, by conducting an experiment in which a naive participant interacted with a confederate. When replicating the experiment with pairs of two naive participants, however, Lockridge and Brennan (2002) were able to show that interlocutors did take into account the common ground after all. Since in our experiments no hearer was present at all, it is unlikely that the strong tendency toward consistency was due to the rather subtle effects of considerations of common ground. It is more plausibe that speakers primed themselves to include attributes previously included and reuse constructions. Whatever the underlying mechanisms are, the finding of such a strong tendency toward consistency has clear implications for the way experimental studies of referential behavior should ideally be designed. Our experiments show that decisions about the design, with respect to the conditions, and the non-critical trials have a significant effect on the results.

The present study has implications for the modeling of referring expression production, as is aimed at in the field of Referring Expression Generation (REG), which is a subfield of computational linguistics. REG models typically consist of an algorithm which generates a referring expression which distinguishes the referent from all other objects in a given context. The output of the algorithms are often evaluated against human-produced referring expressions. It was Pechmann's (1989) study, discussed in Section 2.2.2, which inspired Dale and Reiter (1995) to propose their now classic Incremental Algorithm, which selects attributes incrementally and in a predefined order (a "preference order"). Thus, the algorithm incorporates Pechmann's main finding, namely, that some attributes (such as color) are preferred and therefore selected before others (such as size). The Incremental Algorithm is very influential because it is conceptually and computationally simple, and hence efficient and easy to implement. However, there are several problems with this and related, more recent algorithms (Gatt et al., 2011; Krahmer and van Deemter, 2012).

First, the Incremental Algorithm is under-determined: it does not contain a procedure for finding a preference order (Krahmer and van Deemter, 2012). One way to overcome this problem is to collect production data which indicate what attribute preferences human speakers show when they produce referring expressions. Our study not only shows that color is preferred over pattern and that pattern is preferred over size, but also how salience and absoluteness contribute to those preferences. A second and more important problem is that the Incremental Algorithm is deterministic: in a given situation, it will always produce the same referring expression (Gatt et al., 2011). This is at odds with our finding that there is considerable variation across speakers (see also e.g., Viethen and Dale, 2010). Moreover, the Incremental Algorithm does not take into account the referring expressions that have been produced before in the discourse context. As was discussed in Section 2.3, however, more recent learning models that are able to align with their own previously produced referring expressions have been found to outperform models that do not take into account previously produced referring expressions (Viethen et al., 2014). Importantly, our findings indicate that including one attribute (such as color) can lead speakers to include another attribute (such as pattern), and that *not* including one attribute (such as size) can lead to not including another attribute (such as color and pattern). Modeling this behavior requires a selection procedure that is much more fine-grained than the procedure of the Incremental Algorithm and related algorithms.

Our study indicates that attributes vary in how likely they are to be selected when modification is not necessary. Speakers tend to include color, which is highly salient as well as absolute. The tendency to include pattern is less strong. Since pattern is like color in being absolute, this may suggest that pattern is less salient than color, and that salience is an important factor in the tendency to produce color overspecification, as proposed by Arts et al. (2011a), Gatt et al. (2013), and Koolen et al. (2013). Finally, our study shows that overspecification of size is rare when there is no relevant size contrast in the context, even if size is highly salient. The fact that the presence of a relevant size contrast matters strongly suggests that absoluteness is an important factor in the production of color overspecification, which has been argued before by Pechmann (1989) and Belke and Meyer (2002). However, even size overspecification can be triggered by mentioning the three attributes. In sum, our study indicates that color overspecification is more likely to occur than pattern overspecification because color is more salient than pattern, and much more likely than size overspecification because color is absolute while size is relative.

8. ETHICS APPROVAL

This study was carried out in accordance with the recommendations of the Protocol Ethische Toetsing van Onderzoek (Protocol Ethical Approval of Research), Ethische Toetsingscommissie Geesteswetenschappen (Ethical Committee Faculty of Arts). All subjects gave written informed consent in accordance with the Declaration of Helsinki.

ACKNOWLEDGMENTS

We thank Ronald Fischer for his help in programming the experiments, Nina Kerkhof for testing participants, Roeland van Hout for his help with the statistical analyses, and Emiel Krahmer and two anonymous reviewers for their helpful questions and comments. This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO) to BG, which is gratefully acknowledged.

REFERENCES

Alferink, I., and Gullberg, M. (2014). French-Dutch bilinguals do not maintain obligatory semantic distinctions: evidence from placement verbs. *Bilingualism* 17, 21–37. doi: 10.1017/S136672891300028X

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011a). Overspecification facilitates object identification. *J. Pragmatics* 43, 361–374. doi: 10.1016/j.pragma.2010.07.013

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011b). Overspecification in written instruction. *Linguistics* 49, 555–574. doi: 10.1515/ling.2011.017

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Belke, E. (2006). Visual determinants of preferred adjective order. *Vis. Cogn.* 14, 261–294. doi: 10.1080/13506280500260484

Belke, E., and Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: analyses of viewing patterns and processing times during "same"–"different" decisions. *Eur. J. Cogn. Psychol.* 14, 237–266. doi: 10.1080/09541440143000050

Bock, J. K. (1986a). Meaning, sound, and syntax: lexical priming in sentence production. *J. Exp. Psychol.* 12, 575–586. doi: 10.1037/0278-7393.12.4.575

Bock, J. K. (1986b). Syntactic persistence in language production. *Cogn. Psychol.* 18, 355–387. doi: 10.1016/0010-0285(86)90004-6

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol.* 22, 1482–1493. doi: 10.1037/0278-7393.22.6.1482

Brown, P. M., and Dell, G. S. (1978). Adapting production to comprehension: the explicit mention of instruments. *Cogn. Psychol.* 19, 441–472. doi: 10.1016/0010-0285(87)90015-6

Brown-Schmidt, S., and Konopka, A. E. (2011). Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversations. *Information* 2, 302–326. doi: 10.3390/info2020302

Christ, R. E. (1975). Review and analysis of color coding research for visual displays. *Hum. Factors* 17, 542–570.

Cinque, G. (1994). "On the evidence for partial N-movement in the Romance DP," in *Paths Towards Universal Grammar: Studies in Honor of Richard S. Kayne*, eds G. Cinque, J.-Y. Pollock, L. Rizzi, and R. Zanuttini (Washington, DC: Georgetown University Press), 85–110.

Dale, R., and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902/3

Davies, C., and Katsos, N. (2013). Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al. (2006). *J. Pragmat.* 49, 78–106. doi: 10.1016/j.pragma.2013.01.004

van Deemter, K., Gatt, A., van Gompel, R. P., and Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Top. Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01187.x

Engelhardt, P. E., Bailey, K. G., and Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Engelhardt, P. E., Barış Demiral, Ş., and Ferreira, F. (2011). Over-specified referring expressions impair comprehension: an ERP study. *Brain Cogn.* 77, 304–314. doi: 10.1016/j.bandc.2011.07.004

Ferreira, T. C., and Paraboni, I. (2014). "Referring expression generation: taking speakers' preferences into account," in *Text, Speech, and Dialogue. Proceedings of the 17th International Conference* (Brno), 539–546.

Franke, M. (2012). "Scales, salience and referential language use," in *Logic, Language and Meaning: 18th Amsterdam Colloquium*, eds M. Aloni, F. Roelofsen, G. W. Sassoon, K. Schulz, and M. Westera (Amsterdam; Berlin; Heidelberg: Springer Verlag), 311–320. (December 19–21, 2011, Revised Selected Papers).

Gatt, A. (2007). *Generating Coherent References to Multiple Entities.* Ph.D. dissertation, University of Aberdeen.

Gatt, A., Krahmer, E., van Gompel, R. P., and van Deemter, K. (2013). "Production of referring expressions: preference trumps discrimination," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (Berlin), 483–488.

Gatt, A., van Gompel, R. P., Krahmer, E., and van Deemter, K. (2011). "Non-deterministic attribute selection in reference production," in *Proceedings of the 2nd PRE-Cog Sci Workshop* (Boston, MA).

Gauthier, I., and Tarr, M. J. (1997). Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vis. Res.* 37, 1673–1682. doi: 10.1016/S0042-6989(96)00286-6

Goudbeek, M., and Krahmer, E. (2012). Alignment in interactive reference production: content planning, modifier ordering, and referential overspecification. *Top. Cogn. Sci.* 4, 269–289. doi: 10.1111/j.1756-8765.2012.01186.x

Grainger, J., and Ferrand, L. (1996). Masked orthographic and phonological priming in visual word recognition and naming: cross-task comparisons. *J. Mem. Lang.* 35, 623–647. doi: 10.1006/jmla.1996.0033

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vis. Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Koolen, R., Goudbeek, M., and Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cogn. Sci.* 37, 395–411. doi: 10.1111/cogs.12019

Koolen, R., Krahmer, E., and Swerts, M. (2015). How distractor objects trigger referential overspecification: testing the effects of visual clutter and distractor distance. *Cogn. Sci.* doi: 10.1111/cogs.12297. [Epub ahead of print]. Available online at: http://onlinelibrary.wiley.com/doi/10.1111/cogs.12297/full

Krahmer, E., and van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI/a/00088

Levelt, W. J. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Livingstone, M., and Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science* 240, 740–749. doi: 10.1126/science.3283936

Lockridge, C. B., and Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychon. Bull. Rev.* 9, 550–557. doi: 10.3758/BF03196312

Mangold, R., and Pobel, R. (1988). Informativeness and instrumentality in referential communication. *J. Lang. Soc. Psychol.* 7, 181–191. doi: 10.1177/0261927X8800700403

Meyer, D. E., and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *J. Exp. Psychol.* 90, 227–234. doi: 10.1037/h0031564

Mitchell, M., van Deemter, K., and Reiter, E. (2011). "Applying machine learning to the choice of size modifiers," in *Proceedings of the 2nd PRE-Cog Sci Workshop* (Boston, MA).

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: evidence for facilitatory and inhibitory processes. *Mem. Cogn.* 4, 648–654. doi: 10.3758/BF03213230

Paraboni, I., van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: making referents easy to identify. *Comput. Linguist.* 33, 229–254. doi: 10.1162/coli.2007.33.2.229

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 89–110. doi: 10.1515/ling.1989.27.1.89

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–226. doi: 10.1017/S0140525X04000056

Potter, M. C., and Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *J. Mem. Lang.* 38, 265–282. doi: 10.1006/jmla.1997.2546

Rossion, B., and Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition. *Perception* 33, 217–236. doi: 10.1068/p5117

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguist. Res.* 32, 3–23. doi: 10.1023/A:1021928914454

Sedivy, J. C. (2005). "Evaluating explanations for referential context effects: evidence for Gricean mechanisms in online language interpretation," in *Approaches to Studying World-situated Language Use: Bridging the Language-as-product and Language-as-action Traditions*, eds J. C. Trueswell and M. K. Tanenhaus (Cambridge, MA: MIT Press), 345–364.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., and Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition* 71, 109–147. doi: 10.1016/S0010-0277(99)00025-6

Sonnenschein, S., and Whitehurst, G. J. (1982). The effects of redundant communications on the behavior of listeners: does a picture need a thousand words? *J. Psycholinguist. Res.* 11, 115–125. doi: 10.1007/BF01068215

Sproat, R., and Shih, C. (1991). "The cross-linguistic distribution of adjective ordering restrictions," in *Interdisciplinary Approaches to Language*, eds C. Georgopoulos and R. Ishihara (Dordrecht: Kluwer Academic Publishers), 565–593.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Percept. Psychophys.* 51, 599–606. doi: 10.3758/BF03211656

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Turatto, M., and Galfano, G. (2001). Attentional capture by color without any relevant attentional set. *Percept. Psychophys.* 63, 286–297. doi: 10.3758/BF03194469

Van Der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *J. Mem. Lang.* 60, 448–463. doi: 10.1016/j.jml.2008.12.003

van Gompel, R. P., Gatt, A., Krahmer, E., and van Deemter, K. (2014). "Overspecification in reference: modelling size contrast effects," *Poster Presented at AMLaP 2014* (Edinburgh, UK).

Viethen, J., Dale, R., and Guhe, M. (2014). Referring in dialogue: alignment or construction? *Lang. Cogn. Neurosci.* 29, 950–974. doi: 10.1080/01690965.2013.827224

Viethen, J., and Dale, R. J. (2010). "Speaker-dependent variation in content selection for referring expression generation," in *Proceedings of the 8th Australasian Language Technology Workshop* (Melbourne, VIC), 81–89.

Westerbeek, H. G. W., Koolen, R. M. F., and Maes, A. A. (2014). "On the role of object knowledge in reference production: effects of color typicality on content determination," in *CogSci 2014: Cognitive Science Meets Artificial Intelligence: Human and Artificial Agents in Interactive Contexts,* eds P. Bello, M. Guarini, M. McShane, and B. Scassellati, 1772–1777.

Williams, L. G. (1966). The effect of target specification on objects fixated during visual search. *Percept. Psychophys.* 1, 315–318. doi: 10.3758/BF03215795

Yoon, S. O., and Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *J. Exp. Psychol.* 40, 919–937. doi: 10.1037/a0036161

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 06 July 2015
doi: 10.3389/fpsyg.2015.00935

Stored object knowledge and the production of referring expressions: the case of color typicality

Hans Westerbeek, Ruud Koolen and Alfons Maes*

Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, Netherlands

When speakers describe objects with atypical properties, do they include these properties in their referring expressions, even when that is not strictly required for unique referent identification? Based on previous work, we predict that speakers mention the color of a target object more often when the object is atypically colored, compared to when it is typical. Taking literature from object recognition and visual attention into account, we further hypothesize that this behavior is proportional to the degree to which a color is atypical, and whether color is a highly diagnostic feature in the referred-to object's identity. We investigate these expectations in two language production experiments, in which participants referred to target objects in visual contexts. In Experiment 1, we find a strong effect of color typicality: less typical colors for target objects predict higher proportions of referring expressions that include color. In Experiment 2 we manipulated objects with more complex shapes, for which color is less diagnostic, and we find that the color typicality effect is moderated by color diagnosticity: it is strongest for high-color-diagnostic objects (i.e., objects with a simple shape). These results suggest that the production of atypical color attributes results from a contrast with stored knowledge, an effect which is stronger when color is more central to object identification. Our findings offer evidence for models of reference production that incorporate general object knowledge, in order to be able to capture these effects of typicality on determining the content of referring expressions.

Keywords: reference production, color typicality, content determination, cognitive visual saliency, models of reference production

Edited by:
Kees Van Deemter,
University of Aberdeen, UK

Reviewed by:
Barbara C. Malt,
Lehigh University, USA
Margaret Mitchell,
Microsoft Research, USA

****Correspondence:***
Hans Westerbeek,
Tilburg Center for Cognition and Communication, Tilburg University, Room D412,
P. O. Box 90153, 5000 LE Tilburg, Netherlands
h.g.w.westerbeek@tilburguniversity.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *06 January 2015*
Accepted: *22 June 2015*
Published: *06 July 2015*

Citation:
Westerbeek H, Koolen R and Maes A (2015) Stored object knowledge and the production of referring expressions: the case of color typicality. Front. Psychol. 6:935. doi: 10.3389/fpsyg.2015.00935

Introduction

In everyday language use, speakers often refer to objects by describing what they see, in such a way that an addressee can uniquely identify the intended object (e.g., Pechmann, 1989; Brennan and Clark, 1996; Horton and Gerrig, 2005; Arnold, 2008; Van Deemter et al., 2012a). In **Figure 1**, for example, a speaker can refer to the leftmost object by using the definite description "the yellow tomato." In this visual context this referring expression accommodates unambiguous identification by the addressee, as it describes the target object and rules out the other (distractor) objects. Note, however, that a description like "the tomato" would also suffice as an unambiguous description of the leftmost object, as there are no other tomatoes in the context. Then why would a speaker mention the tomato's color anyway?

A reason could be that the color of the yellow tomato in **Figure 1** draws attention, because it contrasts with one of the features in a stored representation of tomatoes in the speaker's long-term memory, namely the feature that tomatoes are typically red. This makes the color of the tomato

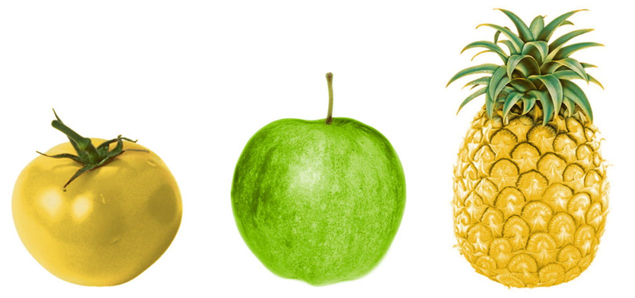


FIGURE 1 | An example of a visual context, containing an atypically colored object. Manipulations of color may not be visible in some print versions of this paper.

cognitively salient. Cognitive salience is different from physical salience, which is visual salience caused by image-level characteristics such as bright colors and strong contrasts (we take the terms cognitive and physical salience from Landragin, 2004). As such, the tomato's color may not be physically different from the color of the pineapple, but when cognitively processed the color of the tomato is more conspicuous. As speakers are inclined to mention object properties that capture their attention or the attention of the addressee (e.g., Krahmer and Van Deemter, 2012), the yellow tomato's atypical color probably causes the speaker to include this in the referring expression, even though this property may not be strictly necessary for unique identification. If speakers are influenced by atypical colors, that implies that speakers are sensitive to contrasts with stored object knowledge when they determine the content of a referring expression.

The question of content determination (i.e., which properties of an object does a speaker include in a referring expression?) is often addressed from both a psycholinguistic perspective and in the field of natural language generation (NLG). Psycholinguistics provides models of content determination by human speakers (e.g., Brennan and Clark, 1996; Engelhardt et al., 2011), for example by addressing the question whether object properties are mentioned merely because they are salient to the speakers themselves, or also because these properties may be found useful for the addressee, whose task it is to identify the referred-to object (e.g., Brennan and Clark, 1996; Horton and Keysar, 1996; Arnold, 2008). NLG models make comparable predictions on content determination, as they often aim to simulate human referring behavior (e.g., Dale and Reiter, 1995; Frank and Goodman, 2012; Krahmer and Van Deemter, 2012).

Models of reference, either implicitly or explicitly, describe at least two (addressee-oriented and speaker-internal) types of factors that speakers rely on when determining the content of a referring expression. The first is how informative an object property is for addressees: when, for example, a property is unique to an object in a context, this property is highly informative with respect to the addressees' task to identify the target object, as it rules out all other objects in the context. As such, informativeness can be regarded as a mainly addressee-oriented factor in content determination. The other factor, salience, is essentially more speaker-internal: speakers tend to mention object properties that capture their visual attention (e.g., Conklin and McDonald, 1982; Brennan and Clark, 1996; Fukumura et al., 2010; Frank and Goodman, 2012; Krahmer and Van Deemter, 2012). This is not to say that addressees would not benefit from object properties that are included in a referring expression based on salience. Speakers' decisions with respect to content determination may reflect addressee-oriented considerations as well (we will further elaborate on this in the general discussion).

While both informativeness for addressees and salience for speakers are part of current models of content determination in reference production, specific extensions may be needed to capture the potential effects of atypicality on content determination. Without such extensions, models of reference would not predict that atypical colors are more salient to speakers (and addressees), and thus would model referring expressions that are identical despite differences in color atypicality.

To test how atypicality may affect content determination, we focus on atypical colors, and study definite descriptions produced by speakers referring to typically and atypically colored objects. Our hypotheses are: (1) A higher proportion of descriptions will include the color of atypically colored objects, compared to typically colored ones; (2) this proportion is correlated to the degree to which a color is atypical for an object; and (3) this proportion is higher when shape is less diagnostic for the identity of an object. Our null hypothesis would be that speakers base content determination on informativeness and physical salience, and thus would not be sensitive to differences in atypicality of target objects.

Theoretical Background

The cognitive processes that underly our predictions for effects of color atypicality on reference production are rooted in the psychology of object recognition. Object recognition is an integral part of speaker-internal processes in reference production. When speakers refer to visually perceived objects, such as the tomato in **Figure 1**, they must first recognize and identify this object as being a member of the category *tomato*. Recognizing objects implies assessing a stored representation of an object in long-term memory, which in turn yields a phonological representation of the object's name (e.g., Humphreys et al., 1988). This will then be realized as the head noun of the referring expression. Stored knowledge of the typical colors of objects plays a role in this process of object recognition and naming.

That atypicality affects object recognition follows from work in experimental psychology (e.g., Tanaka and Presnell, 1999; Tanaka et al., 2001; Therriault et al., 2009). In several studies, it is shown that color plays a role in object recognition through response latencies for example, as people are slower to recognize and name objects that are atypically colored (e.g., Price and Humphreys, 1989; Therriault et al., 2009), or through Stroop tasks (Naor-Raz et al., 2003). These effects are caused by the fact that an atypical color cannot function as a useful cue for finding the corresponding mental representation of the object. Also, atypically colored objects are visually salient and thus likely attract attention in a scene (e.g., Becker et al., 2007). These studies show that for (at least some) objects color is part of an object's representation in stored knowledge, and that this is accessed when objects are recognized (see Tanaka et al., 2001 and Bramão et al., 2011a, for comprehensive reviews).

Not all objects are strongly tied to one or a few particular colors. The degree to which a particular object is associated with a specific color is called color diagnosticity (e.g., Tanaka and Presnell, 1999). Objects that can have any color are called non-color-diagnostic. The color of these objects is not predictable from the object's category (e.g., Sedivy, 2003; Bramão et al., 2011a), as theys can have many different colors (e.g., cars, pens). Conversely, objects that do have one or a few prototypical colors associated with them are called color-diagnostic objects (e.g., bananas, carrots), because color is diagnostic in determining their identity, and can be predicted from the object's category (e.g., Tanaka and Presnell, 1999; Bramão et al., 2011a,b).

To study effects of atypicality, the focus is on color-diagnostic objects, because the color of these objects can be more or less like the prototypical color of the category the object belongs to. As said, in stored knowledge, the mental representation of such objects plausibly contains information about what their typical color is (e.g., Naor-Raz et al., 2003). This information is based on the color of objects in the same ontological category: if many exemplars of an object have the same color, then this color is prototypical of the object's category (e.g., Rosch and Mervis, 1975). This does not rule out that other colors are possible too: Rosch's (1975) Prototype Theory postulates that one object exemplar can simply be a better representative of the category than another. So, the exact color used is one factor that determines how atypical a color is for an object: for example, blue is very atypical for bananas, but green not so much.

Within the category of color-diagnostic objects, higher, and lower color-diagnostic objects can be distinguished (e.g., Tanaka and Presnell, 1999). For high color-diagnostic objects, color is an important feature in determining their identity. Typical examples of such objects are fruits: often a fruit's shape is simple and similar to other fruits (i.e., round with only a few protruding parts), which makes color more diagnostic in identification (e.g., Tanaka et al., 2001). So, when other aspects of objects such as shape are more characteristic, color is likely to be less instrumental in object recognition (Rosch and Mervis, 1975; Mapelli and Behrmann, 1997; McRae et al., 2005; Bramão et al., 2011a, p. 245). Shape diagnosticity is, for object recognition, a moderating factor in the degree of association between an object and its typical and atypical colors: once viewers have to recognize atypically colored objects having a highly diagnostic shape, we may expect color to be less crucial in the recognition of the object, as the process will be informed more prominently by the diagnostic shape. It may be assumed that manipulations of color typicality are more conspicuous for objects with a relatively simple shape (e.g., lemons) than for complex-shaped objects (e.g., lobsters).

As color atypicality is important for object recognition (and more so if objects have a low-diagnostic shape), and atypical colors capture visual attention (Landragin, 2004; Becker et al., 2007), what does that mean when speakers have to produce an adequate referential expression for visually present objects? In general, speakers are inclined to mention what captures their visual attention in referring expressions, which may be useful for addressees (e.g., Conklin and McDonald, 1982; Brennan and Clark, 1996; Keysar et al., 1998; Fukumura et al., 2010; Frank and Goodman, 2012; Krahmer and Van Deemter, 2012). Hence, for physical salience, the link with content determination is indeed well-established. For example, color contrast causes speakers to mention color in their object descriptions (e.g., Viethen et al., 2012; Koolen et al., 2013). But what about cognitive salience, and color (a)typicality in particular? We expect that the cognitive salience associated with atypical colors also results in color being a highly preferred attribute when speakers have to produce adequate referential expressions for atypically colored objects.

The idea that stored knowledge of typical colors of objects plays a role in content determination gains support from a production study by Sedivy (2003). Her work does not involve atypical colors, but she investigated whether speakers mention color in a referring expression dependent on the color diagnosticity of the objects they describe. Participants gave instructions to a conversational partner to move one of two (typically) colored drawings of objects. In the experimental trials, color was not necessary for helping the addressee to disambiguate the target object from the other object, so mentioning color would yield what is called an overspecified referring expression (e.g., Pechmann, 1989; Koolen et al., 2011). The target objects (i.e., those that were to be moved) were either color-diagnostic (e.g., yellow bananas), or non-color-diagnostic (e.g., yellow cars). Sedivy (2003) observed that for color-diagnostic objects, the proportion of speakers that mentioned the (predictable) color of such objects was roughly thirty percent lower than when objects were not color-diagnostic. All objects in Sedivy's experiment were typically colored, and it is yet unclear whether colors that contrast with stored knowledge will also make speakers include color. Sedivy's (2003) results, however, do suggest that content determination is affected by color information in object knowledge, and that speaker's decisions to encode color in a referring expression are not taken independently of an object's type.

Participants in a study by Mitchell et al. (2013a) described objects with atypical materials or shapes, where mentioning these properties was necessary for the addressee to uniquely identify the intended object. Although not dealing with color, Mitchell et al.'s (2013a) study directly suggests that atypical object properties are preferred over typical ones in content determination. In their experiment, participants instructed a lab assistant to move a number of objects on a table into positions in a grid. Target objects could not be uniquely identified by mentioning their type only, so participants had to include shape, texture, or both in their referring expressions in order to be unambiguous. Crucially, Mitchell et al. (2013a) manipulated whether the shape of the object was atypical (e.g., a hexagonal mug), or whether the material was atypical (e.g., a wooden key), and using neither of those properties would result in an ambiguous referring expression. Thus, for unique identification of the target objects the speakers had to decide between mentioning a typical property, an atypical one, or both. Speakers turned out to prefer the atypical property over the typical one significantly more often than the other way around.

So, previous work on reference production in combination with color diagnosticity and typicality shows that speakers to mention atypical properties of objects when referring to them. Nonetheless, there are some ways in which this work can be extended, with respect to overspecification, effects of color diagnosticity and typicality in object recognition, and the specific use of color adjectives. Firstly, it is yet unclear whether atypicality leads speakers to mention an atypical property that is not needed to uniquely identify the target object, but will yield an overspecified referring expression instead. In Mitchell et al.'s (2013a) task, mentioning the atypical property always disambiguated the target object from distractors, and as such one can speculate that the preference of speakers for the atypical property over the typical one may not only be due to the atypicality *per se*, but also because speakers may have found the atypical property somehow more informative or useful than the typical alternative. Such decisions may be different when the atypical property is not needed to uniquely identify the object. Secondly, Mitchell et al.'s (2013a) data does not provide insight into a potential relationship between the degree of atypicality of an object property and the probability that it is included in a referring expression. It may be less straightforward to define a degree of atypicality for a shape or material given some object, but this is possible in the case of color typicality. Finally, we argue that it is interesting to look specifically at color, because color is often found to be one of the most salient properties of objects and is realized in referring expressions more often than any other property (e.g., Pechmann, 1989), also in more naturalistic domains (e.g., Mitchell et al., 2013b).

The Current Experiments

To investigate how effects of color atypicality in object recognition may affect content determination in reference production, we test whether speakers redundantly include color in a referring expression, and whether this is proportional to the degree of (a)typicality of that color for the object that is referred to. Following the object recognition literature, the degree to which specific objects are associated with particular colors theoretically depends on two factors. One factor is the degree of color atypicality: Some colors are more atypical for an object than other colors (e.g., blue bananas are more atypical than green ones). The other factor is shape diagnosticity: manipulations of color typicality are expected to be more conspicuous for low-shape-diagnostic objects (e.g., lemons) than for high-shape diagnostic ones (e.g., lobsters), because for the former type of objects color may be less crucial in object recognition. Given the integral role of object recognition in reference production, the question is how these factors affect the production of referring expressions.

In two language production experiments, speakers view simple visual contexts comprised of multiple typically and atypically colored objects. The speakers are instructed to describe one of the objects in such a way that a conversational partner can uniquely identify this target object. The contexts are constructed as such that color is never necessary for unique identification. As such, we keep the informativeness of color for the addressees' task to identify the intended referent equal across all conditions. So, when speakers mention color, this is in a strict sense redundant. In Experiment 1, we investigate how the degree of atypicality of a color for the target object (on a continuum, established in a pretest) affects the proportion of descriptions including color. We aim to maximize the diagnostic value of color by focusing on objects with a low-diagnostic shape (e.g., Bramão et al., 2011a). In Experiment 2, we compare typically and atypically colored objects that have a shape that is more versus less diagnostic, in order to address the second factor that is expected to moderate color typicality. So, we investigate whether our findings from the first experiment extend to objects for which color itself is a less central property, and whether shape diagnosticity moderates speaker's sensitivity to color atypicality in reference production.

Experiment 1: Referring to Objects with Colors of Different Degrees of Atypicality

Method

Participants

Forty-three undergraduates (eleven men, thirty-two women, median age 21 years, range 18-25) participated for course credit. The participants were native speakers of Dutch (the language of the study). All gave consent to have their voice recorded during the experiment. Their participation was approved by the ethical committee of our department.

Materials Pretest

A pretest was conducted to determine the degree of atypicality of objects in certain colors. Sixteen high-color-diagnostic objects were selected on the basis of stimuli used in object recognition studies (e.g., Naor-Raz et al., 2003; Therriault et al., 2009). These objects were mainly fruits and vegetables, with simple shapes. In terms of geons (cf., Biederman, 1987), they were mainly comprised of one or two simple geometric components. Such simple objects have an uncharacteristic shape, as shape is relatively uninformative for distinguishing these objects from other object categories (Tanaka et al., 2001). This makes color more instrumental in object recognition (Bramão et al., 2011a). For each of the objects a high quality photograph was obtained, which was edited such that the object was on a plain white background. Further photo editing was done to make a red, blue, yellow, green, and orange version of each object. This resulted in a set of eighty photos (16 object types in five colors).

The photos were presented to forty participants in an on-line judgment task (thirteen men, twenty-seven women, median age 22.5 years, range 19-54; none participated in any of the other experiments and pretests in this paper). To manage the length of this task, participants were randomly assigned to one of two halves of the photo set. For each photo, participants first had to type in the name of the object ("what object do you see above?") and the object's color ("which color has the object?"). Then, they answered the question "how characteristic is this color for this object?" by using a slider control ranging from "is not characteristic" to "is characteristic" ("niet kenmerkend," "wel kenmerkend" in Dutch). The position of the slider was linearly converted to a typicality score ranging from 0 to 100, where 100 indicated that the color-object combination was judged as most typical (i.e., the slider was placed in the rightmost position). For each photograph, the typicality score was averaged over participants in order to calculate a measure of color typicality.

Materials

Based on the results of the pretest, fourteen objects were selected for the experiment. Two objects were rejected because typicality scores were low for all the colors tested, or because many participants had difficulties naming the object (see the supplementary materials for details). Furthermore, of each object two colors were discarded, such that the final set of objects and colors would represent the whole spectrum of the typicality ratings continuum obtained in the pretest (scores ranging from 2 to 98, from very atypical to very typical, plus scores in between). As an illustration: the least typical objects were a blue bell pepper and red lettuce, among the most typical ones were yellow cheese and a red tomato. A yellow apple and a green tomato fell about halfway in between the extremes.

The final set of objects was used to construct forty-two experimental visual contexts. **Figure 2** presents three examples of these contexts. Each context contained six different objects, positioned randomly in a three by two grid. The colors of these objects were chosen such that there were three different colors in each context, with each color appearing on two objects. Also, the typicality score averaged over the six objects in each context was similar for all trials (the mean typicality score of each context was between 40 and 60). One of the objects in each context was the target object, which was marked with a black square outline. The other five objects were the distractors. The target object was always of a unique type in each context, so mentioning the target object's color was never necessary to disambiguate the target from any of the distractors. Crucially, the 42 target objects differed in their degree of typicality, as established in the pretest.

To ensure that the degree of color typicality of the target object was not confounded with physical salience, we assessed salience by using a computational perceptual salience estimation algorithm (Erdem and Erdem, 2013). We did this because any effect of color atypicality on whether speakers mention color in a referring expression should not be attributable to the object's color being more bright, contrasting, or otherwise physically salient to the speaker. Crucially, the algorithm that we used does not incorporate any general knowledge about objects and their typical colors, as it only measures salience based on physical (image-level) features.

We ran Erdem and Erdem's (2013) algorithm on our 42 experimental visual contexts, using its standard settings and parameters. The algorithm outputs physical salience scores for each pixel of an image, which expresses the relative salience of that pixel with respect to other pixels in the image. In our visual contexts, six areas of interest (AOIs) were defined, one for the target object and five for the distractor objects. Of each AOI, the mean relative salience of the pixels was calculated, which expresses how salient the object in that AOI is compared to the other AOIs (i.e., objects) in the context.

Analyses of the mean relative salience as determined by the algorithm showed that there was no significant correlation between the degree of physical salience of the target object in each scene and its color typicality, Pearson $r(40) = 0.05$, $p = 0.721$. The atypically colored objects in our experiment were physically not more salient than the typically colored ones (and vice versa). Furthermore, a one-way analysis of variance with color as the independent and salience as the dependent variable showed no differences in salience for each of the five target colors, $F(4,41) = 1.05$, $p = 0.397$.

In addition to the experimental contexts, we created 42 filler contexts. These consisted of four hard-to-describe greebles (Gauthier and Tarr, 1997), all purple, so that participants were not primed with using color in the other trials. One greeble was marked as the target object that had to be distinguished from the distractors.

Procedure

Participants sat at a table facing the experimenter, with a laptop in front of them. The participants were presented with the 42 trials, one by one, on the laptop's screen. Between each experimental trial, there was a filler trial. Participants

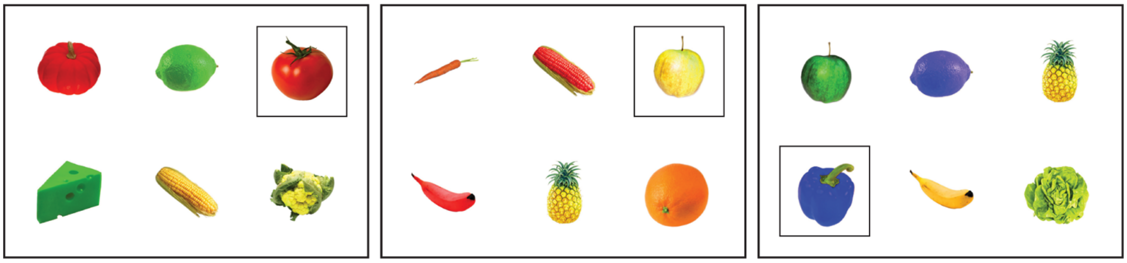


FIGURE 2 | Examples of visual contexts in Experiment 1. From left to right: a context with a highly typical target (red tomato; typicality score 97), one with a not typical nor atypical target (yellow apple; typicality score 58), and one with an atypical target (blue pepper; typicality score 2).

described the target objects in such a way that the experimenter would be able to uniquely identify them in a paper booklet. The instructions emphasized that it would not make sense to include location information in the descriptions, as the addressee would see the objects in a different configuration. Participants could take as much time as needed to describe the target, and their descriptions were recorded with a microphone. The addressee (experimenter) never asked the participants for clarification, so the data presented here are one-shot references.

The procedure commenced with two practice trials: one with six non-color-diagnostic objects in different colors, and one practice trial with greebles. Once the target was identified, this was communicated to the participant, and the experimented pressed a button to advance to the next trial. The trials were presented in a fixed random order (with one filler after each experimental trial). This order was reversed for half of the participants, to counterbalance any potential order effects. After completion of the experiment, none of the participants indicated that they had been aware of the goal of the study. The experiment had an average running time of about 25 min.

Research Design and Data Analysis

For each of the experimental trials, we determined whether the speakers' description of the target object resulted in unambiguous reference, which mainly implied annotating whether respondents used the correct type attribute. Because the target object was always of a unique type in each context, mentioning type was sufficient. We also assessed whether the object's type was named correctly. Using the correct type was important, because otherwise we could not deduce whether the object's color was regarded as typical or atypical. We annotated each description as either containing a color adjective, or not.

Whether mentioning color was related to the degree of color atypicality of the target object was analyzed using logit mixed models (Jaeger, 2008). Initial analyses revealed that stimulus order had no effects, so this was left out in the following analyses. In our model, color typicality (as scores on the pretest) was included as a fixed factor, standardized to reduce collinearity and to increase comparability with Experiment 2. Participants and target object types were included as random factors. The model had a maximal random effect structure: random intercepts and random slopes were included for all within-participant and within-item factors, to ensure optimal generalizability (Barr et al., 2013). Specifically, the model contained random intercepts for participants and target objects, and a random slope for color typicality at the participant level.

Results and Discussion

The data of three participants was not analyzed because of technical issues with the audio recordings. Of the remaining 1680 descriptions, 1629 descriptions (97%) were intelligible, unambiguous, and contained a correct type attribute, resulting in unique reference. As expected, practically all analyzed descriptions were of the form "the tomato" or "the yellow tomato."

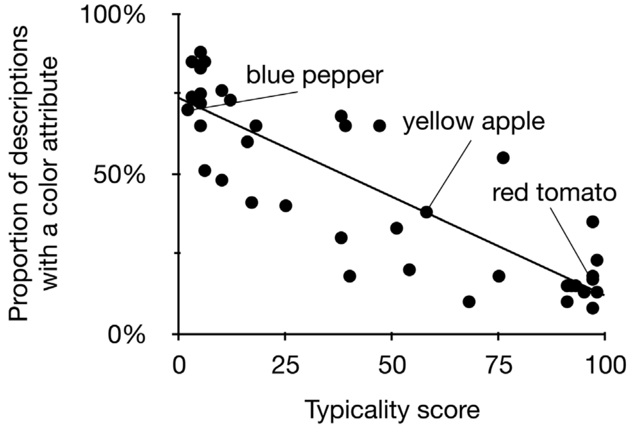


FIGURE 3 | Typicality scores of objects (horizontal axis) and the proportion of descriptions of these objects that contain color (vertical axis) in Experiment 1. Some illustrative objects are labeled in this plot; the line represents the correlation between the two variables.

Figure 3 plots the atypicality score of a target object in the pretest against the proportion of descriptions that contained color in the production experiment (exact proportions and typicality scores are listed in the Supplementary Materials). The mixed model revealed a significant effect of color typicality on whether a target description contained a color attribute or not, $\beta = -2.36$, $SE = 0.25$, $p < 0.001$. The direction of the effect indicated that lower typicality in the pretest was associated with more referring expressions containing color. An additional analysis by means of bivariate correlation between the typicality score of each object and the proportion of speakers mentioning color for this object reconfirmed that these were significantly related, Pearson $r(40) = -0.86, p < 0.001$.

The results of our experiment warrant the conclusion that content determination is affected by the degree of typicality of a target object's color. When a color is more atypical for an object, the proportion of referring expressions that include that property increases. This effect is very strong, as exemplified by the high correlation between the two variables. **Figure 3** also suggests that it is highly consistent across speakers: for a considerable number of typically colored stimuli, the percentage of speakers not using color approaches zero, and conversely, for some atypically colored stimuli this percentage approaches 100%. This supports the theory that speakers evaluate contrasts with stored knowledge about typical features of objects in long term memory when producing a referring expression.

In Experiment 1, we have manipulated the degree of atypicality of the target objects by using different colors for objects, such that the object-color combinations span a range of atypicality scores. For example, speakers have described blue tomatoes (very atypical), green tomatoes (not atypical nor typical), and red tomatoes (very typical). However, target objects in Experiment 1 were predominantly simply shaped fruits and vegetables, i.e., objects for which color is especially instrumental in their identification (as their shape is not very informative about the identity of the objects; Tanaka and Presnell, 1999; Bramão et al., 2011a). As explained in the theoretical background, the diagnostic value of an object's color in recognition is lower when its shape is more diagnostic (Bramão et al., 2011a). Accordingly, would color atypicality be less conspicuous when shape is more diagnostic, resulting in a moderation of the color atypicality effect on reference production? Therefore, the goal of Experiment 2 is to investigate the effect of color typicality on reference production, as a function of objects' shape diagnosticity.

Experiment 2: Referring to Typically and Atypically Colored Objects with High or Low Shape Diagnosticity

In Experiment 2, we cross color typicality with shape diagnosticity in a language production task similar to the one used in Experiment 1. As such, we aim to extend our findings from the first experiment to low-color-diagnostic objects (with more diagnostic shapes). We expect to find a similar relationship between color typicality and content determination as in Experiment 1, but because for low-color-diagnostic objects color is less instrumental in their identification we predict that higher shape diagnosticity overall decreases the proportion of referring expressions that include color. Secondly, we predict that shape diagnosticity and color typicality interact, such that effects of color typicality are larger when shapes are less diagnostic compared to when shapes are more diagnostic.

Method

Participants

Sixty-two undergraduates participated for course credit. They participated in dyads, with one participant acting as the speaker and the other as addressee. So, there were 31 speakers (7 men, 24 women, median age 22 years, range 18-25), all were native speakers of Dutch (the language of the study). None of the participants took part in any of the other experiments and pretests in this paper. They gave consent to have their voice recorded during the experiment. Their participation was approved by the ethical committee of our department.

Materials

High quality white-background photos of 16 target objects were selected and edited, similar to Experiment 1, and supplemented by stimuli used in object recognition studies. The typical color of these objects was either red, green, yellow, or orange. Even though the saliency algorithm we employed showed no differences in physical salience between the five target colors used in Experiment 1, we decided for Experiment 2 to not use blue objects (which were all atypical in Experiment 1), and to equally balance color frequencies throughout the experiment. As such, the proportions of target objects in each color was kept identical in all conditions.

Half of the objects were low in shape diagnosticity: they had relatively simple shapes, as they were mostly round with very few protruding parts, like in Experiment 1. The other objects were high in shape diagnosticity, having relatively complex shapes, comprising many protruding parts and no basic round shape (i.e., comprised of many geons). Such objects (e.g., lobster; see the supplementary materials for a complete list of objects used) thus have a more characteristic (diagnostic) shape, which sets it apart from other object categories.

As in Experiment 1, the target objects were placed in visual contexts of six objects. Again, the colors of these objects were chosen such that there were three different colors in each context, with each color appearing on two objects. Three of the objects were typically colored, the other three atypically colored. One of the objects in each context was the target object, singled out by a black square outline for the speaker. The other five objects were the distractors. The target object was always of a unique type, so that mentioning the target object's color was never necessary to disambiguate the target from any of the distractors.

Eight contexts contained objects that were low in shape diagnosticity, and the other eight contexts contained objects high in shape diagnosticity. Also, in half of the contexts the target object was typically colored, and in the other half it was atypically colored. **Figure 4** presents examples of the contexts in each of the four resulting conditions: the contexts on the left contain a typically colored target object; in the contexts on the right the target has an atypical color. The upper contexts comprised of low shape diagnostic objects; the lower contexts has high shape diagnosticity.

The target objects were subjected to an on-line judgment task similar to the pretest in Experiment 1. Sixteen participants took part in this task (6 men, 10 women, median age 21 years, range 18-26; none participated in any of the other experiments and pretests in this paper). As expected, typically colored objects yielded a higher typicality score (range 87.50-99.75) than atypically colored objects range 0.83-10.50). There were no differences in typicality scores for object with a high and a low shape diagnosticity ($F < 1$), and the two factors did not interact ($F < 1$). The pretest also showed that none of the objects were difficult to name.

As in Experiment 1, we used the computational physical salience estimation of Erdem and Erdem (2013) to ensure that color typicality was not confounded with differences in relative physical salience between typical and atypical objects, and between objects with high and low shape diagnosticity. Analyses of variance of the mean relative salience of the target objects showed no differences between typically colored and atypically colored target objects ($F < 1$), nor between objects with high and low shape diagnosticity ($F < 1$). The two factors did not interact ($F < 1$). This shows that possible (interaction) effects involving shape diagnosticity cannot be ascribed to colors being physically more salient when for example shapes are simple and colored areas may appear to be larger.

Procedure

Participants took part in pairs. Who was going to act as the speaker and who as the addressee was decided by rolling a dice. In contrast to Experiment 1, addressees were naive participants instead of a confederate, in order to improve ecological validity (cf. Kuhlen and Brennan, 2013). Participants were seated opposite

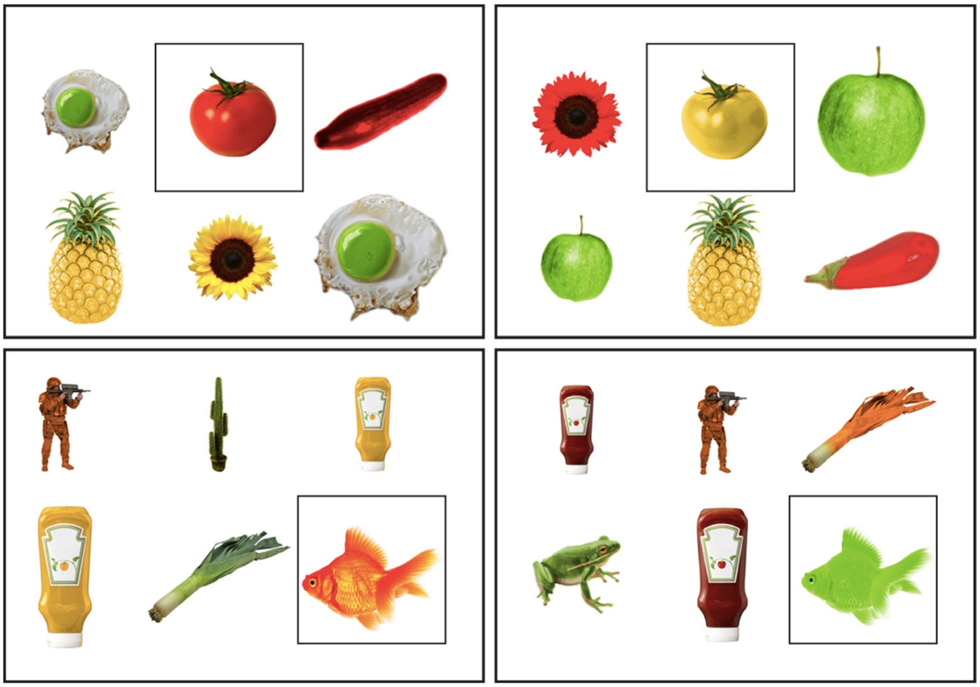


FIGURE 4 | Examples of visual contexts in each of the conditions in Experiment 2, in two color typicality conditions (horizontal axis) and in two shape diagnosticity conditions (vertical axis).

each other at a table, and each had their own computer screen. The screens were positioned in such a way that the face of either participant was not obstructed (ensuring that eye contact was possible), while participants could not see each other's screen.

Each speaker described the target object of the sixteen visual contexts, as well as 32 filler contexts containing purple greebles. We made two lists containing the same critical trials, but with reversed typicality: target objects that were typically colored for one speaker were atypically colored for another. As such, color typicality and shape diagnosticity were manipulated within participants, while ensuring that each target object appeared in only one typicality condition for each participant. We did this because one could speculate that the overall proportion of color adjectives in Experiment 1 might inflate because participants used them to express contrasts between objects of the same type over trials. The order of the contexts in each list was randomized for each participant, but there were always two filler trials between experimental ones (i.e., one more than in Experiment 1, to further assure that that the colorful nature of our stimuli does not boost the overall probability that color was mentioned; see Koolen et al., 2013).

The addressee was presented with the same contexts as the speaker, but without any marking of the target object. Also, the objects on the addressee's screen were in a different spatial configuration than on the speaker's screen, in line with the instruction that it would not make sense for the speaker to mention location information. In each trial, the addressee marked the picture that he or she thought the speaker was describing on an answering sheet. Although the addressee was instructed that clarifications could be asked, there were no such requests during the whole experiment, so the data presented here are one-shot references.

The procedure commenced with two practice trials with greebles, plus one practice trial with non-color-diagnostic objects (as in Experiment 1). Once the addressee had identified a target, this was communicated to the speaker, and a button was pressed to advance to the next trial. The experiment finished when all trials were described and the addressee identified the last target object. The experiment had an average running time of about 15 min.

Research Design and Data Analysis

Data annotation was identical to Experiment 1. We analyzed whether using a color adjective or not was related to the degree of color atypicality of the target object using logit mixed models (Jaeger, 2008). Initial analyses revealed that stimulus list and stimulus order (trial number) had no effects, so these factors were left out in the following analyses. In our model, color atypicality and shape diagnosticity were included as fixed binomial factors, standardized to reduce collinearity and to increase comparability with Experiment 1. Participants and target object types were included as random factors. The model had a maximal random effect structure: random intercepts and random slopes were included for all within-participant and within-item factors, to ensure optimal generalizability (Barr et al., 2013). Specifically, the model contained random intercepts for participants and target objects, random slopes for color atypicality and shape diagnosticity at the participant level, and a random slope for color atypicality at the target object level.

Results and Discussion

In total, 496 target descriptions were recorded in the experiment. 472 descriptions (95%) were intelligible, unambiguous, and contained a correct type attribute, resulting in unique reference. Practically all analyzed descriptions were of the same form as those in Experiment 1.

Our model revealed a significant effect of color atypicality on whether a target description contained a color attribute or not, $\beta = 3.53$, $SE = 0.39$, $p < 0.001$. Of the references to atypically colored target objects, 75.3% contained color, compared to 14.3% for typically colored target objects. Also, the model showed a significant main effect of shape diagnosticity, $\beta = -0.89$, $SE = 0.35$, $p = 0.010$. References to objects with a high diagnostic (i.e., complex) shape contained color in 38.4% of the cases, compared to 49.1% for low diagnostic (i.e., simple) shape target objects. Color typicality and shape diagnosticity interacted, such that the effect of typicality on using color in a referring expression was larger for low shape diagnostic objects than for the high shape diagnostic objects, $\beta = -0.70$, $SE = 0.32$, $p = 0.030$. **Figure 5** plots the proportion of referring expressions containing color for each of the four conditions in the experiment.

With respect to the effect of color typicality on content determination, inspection of the data revealed that not a single speaker acted against the general pattern and mentioned color more often for typically colored objects than for atypically colored ones. However, a mere three speakers mentioned color in all atypical trials, and never mentioned color in the typical trials. While most speakers showed more variation in their response to color atypicality, only these three speakers show what is often called *deterministic behavior* in the literature (e.g., Van Deemter et al., 2012b).

Experiment 2 shows that the effect of color typicality on content determination is moderated by the diagnosticity of an object's shape. Color is more often mentioned for objects with low shape diagnosticity. It is for these objects that the color atypicality effect is slightly larger compared to objects with higher shape diagnosticity. This further supports the idea that object recognition and the status of features of objects in long-term memory is closely related to reference production.

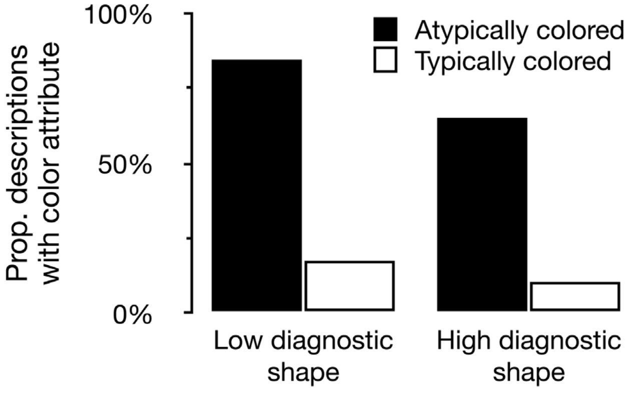


FIGURE 5 | The proportion of referring expressions containing color for each of the four conditions in Experiment 2.

General Discussion

We investigated the role of speakers' stored knowledge about objects when producing referring expression. The experiments reported in this paper show a strong effect of color atypicality on the object properties mentioned by speakers. Speakers mention the color of atypically colored objects significantly more often than when objects are typically colored, and this effect is moderated by the degree of atypicality of the color, and the diagnosticity of the object's shape. These results support the view that stored knowledge about referred-to objects influences content determination. When a property of an encountered object contrasts with this knowledge, the probability that this property is included in a referring expression increases significantly. This also suggests that because object recognition is an integral part of reference production, there may be a close relation between findings in object recognition related to color diagnosticity and typicality on the one hand, and effects on reference production on the other.

Combined with the findings of Mitchell et al. (2013a), who report similar effects of atypical materials and atypical shapes on content determination, the current paper forms converging evidence for sizable effects of atypicality on the production of referring expressions. Furthermore, our results corroborate Sedivy's (2003) finding that object knowledge affects content determination, and that speakers' decisions to encode color in a referring expression are not taken independently of the object's type. Our research also resonates with Viethen et al.'s (2012) findings on how the specific color of an object can affect a speaker's decision to include this color in a referring expression. While Viethen et al.'s (2012). focus on colors that are relatively easy to name or not (e.g., blue versus light blue), we report effects of specific colors combined with specific object types.

We attribute the effects of color atypicality on content determination reported in this paper to the speakers' visual attention allocation, and cognitive salience in particular: because atypical colors attract visual attention (e.g., Becker et al., 2007), speakers tend to encode these colors in a referring expression (e.g., Krahmer and Van Deemter, 2012). In the visual contexts that we used, mentioning the type of the object was always sufficient to fully disambiguate the target object from all the distractors. The speakers' decision to include color is in that sense redundant (i.e., the referring expressions containing color are overspecified; cf. Pechmann, 1989; Koolen et al., 2011). Instead of carefully assessing the objects and their properties in the visual context, and calculating their informativeness, speakers in our experiments appeared to use other rules or mechanisms to determine the content of a referring expression.

The idea that speakers may rely on different content determination processes than calculations of informativeness has been postulated in a number of recent papers (e.g., Dale and

Viethen, 2009; Van Deemter et al., 2012b; Viethen et al., 2012, 2014; Koolen et al., 2013). Instead of a careful consideration of the properties and salience of all (or a subset of) the objects in a visual context, speakers may turn to quicker, simple decision rules to make judgments in the content determination process. Such a decision rule that would fit our data would be: "If the contrast between the color of the target object and stored knowledge is strong, increase the probability that it is mentioned."

Speakers' reliance on relatively simple decision rules is argued to be related to the visual complexity of the contexts that they are confronted with. Some researchers hypothesize that speakers may especially rely on the "fast and frugal heuristics" in cases where considering all properties of all objects in a context is cognitively costly (e.g., Van Deemter et al., 2012b, p. 179). However, the contexts in our experiments are undoubtedly very simple: speakers only have to consider the type of six objects that are presented in an uncluttered and simple environment, which is a task that is arguably well within the speakers information processing capacity (e.g., Miller, 1956). Yet speakers seem to apply (a variation of) the aforementioned decision rule in contexts with an atypically colored target. Such contexts are not more complex or visually cluttered than the typical ones. So, the decision rule that we propose above would not be one that merely applies when the (limited) processing capacity of speakers is exceeded, but one that is universally available whenever the content of a referring expression is determined.

Implications for (Computational) Models of Reference Production

Being able to refer to objects in a human-like manner is an important goal for NLG models of reference production (REG algorithms), and for the field of NLG (a subfield of Artificial Intelligence) in general (Dale and Viethen, 2009; Frank and Goodman, 2012; Van Deemter et al., 2012b). Our findings pose a new challenge for current REG algorithms. In the light of our findings, models can be enhanced by incorporating general object knowledge, because without access to such information they are unable to distinguish between typical and atypical objects when determining the content of a referring expression. Moreover, in our data, the decision to include color in a referring expression appears not to be taken independently of the target object's type. For example, speakers decide to mention redness when they describe a lemon, but not when they describe a tomato. This is something that a model should be able to take into consideration.

Popular NLG models predict color use irrespective of the typicality and diagnosticity of the target's color. In the Incremental Algorithm (IA; Dale and Reiter, 1995), attributes like color, size, and orientation are included in a referring expression on the basis of how informative they are, and they are considered one by one (i.e., incrementally). More salient attributes, like color, are considered early, because they are highly ranked in a predefined preference order (which is typically determined on the basis of empirical data). Type is likely to be included anyway, because it is necessary to create a proper noun phrase, and this would yield fully distinguishing referring expressions in all conditions in our experiments. The IA would therefore generate no color adjectives. If the IA was to be able to make the decision to mention the color of a yellow tomato, for example, and not for a red tomato, it would need a ranking (preference order) of certain colors for tomatoes (e.g., red, green, orange, yellow, blue), instead of a mere ranking of certain attributes (e.g., color, size, orientation).

The model of pragmatic reasoning by Frank and Goodman (2012) allows salience of objects to be modeled for each visual context individually (instead of in a predefined preference order). So, in effect, the salience of atypically colored objects can be modeled to be different from the salience of typically colored ones. However, Frank and Goodman (2012) calculate this (prior) salience on the basis of empirical findings, so behavioral data is needed before reference production is modeled. And while it is well possible to estimate visual salience computationally and automatically (e.g., Erdem and Erdem, 2013), such salience estimations are not (yet) able to take general knowledge into account and thus respond differently to various degrees of atypicality.

The challenge is to feed such salience estimations with knowledge about what prototypical colors of objects are, and how important color is in the identity of these objects. Assuming that object types are readily recognized computationally in a visual context (which works quite well in controlled environments nowadays, Andreopoulos and Tsotsos, 2013), a knowledge base containing prototypical object information can be queried at runtime when a referring expression is generated. This is what Mitchell et al. (2013a) and Mitchell (2013) propose in their discussion of repercussions of atypicality for REG. However, for color, a simpler system without a dedicated knowledge base may be effective too. A web search for images (e.g., on Google Images) may inform an algorithm about color typicality: when the dominant color of the first n image results of a web search is computationally determined, the prototypical color of an object should be derivable. In fact, we expect that this method can even generate the degree of atypicality of a color, much alike the typicality scores that we obtained in a pretest for Experiment 1. A comparison between the n search results showing one color and the n results showing other colors probably yields a good estimation of the degree of atypicality of that particular color.

Our results are also interesting in the light of an observed tendency toward using more naturalistic stimuli in behavioral experiments that are aimed at evaluating computational models of reference production (e.g., Coco and Keller, 2012; Viethen et al., 2012; Clarke et al., 2013; Mitchell, 2013; Mitchell et al., 2013a,b; Koolen et al., 2014). Color typicality may be an important difference between artificial and more naturalistic stimuli, as studies that employ artificial contexts often present speakers with atypically colored objects (e.g., green television sets and blue penguins; Koolen et al., 2013; Viethen et al., 2014). Our results seem to argue against using artificial contexts in reference production studies by showing that content determination can be steadily affected by atypical colors.

Color Atypicality and Speaker-Addressee Perspectives in Reference Production

In our experiments, speakers produced referring expressions for an addressee who was present in the communicative setting.

Although speakers in our experiments presumably mention the color of atypically colored target objects because atypical colors are cognitively salient to the speakers themselves, this does not necessarily assert that mentioning atypical colors more often than typical ones is exclusively speaker-internal behavior (e.g., Keysar et al., 1998; Wardlow Lane et al., 2006; Arnold, 2008). Speakers' decisions to include color may as well be addressee-oriented and reflect what is called *audience design* in the literature (e.g., Clark, 1996; Horton and Keysar, 1996; Arnold, 2008; Fukumura and van Gompel, 2012). As suggested in the general introduction, if speakers take the addressee's perspective into account and use their own perception as a proxy for the addressees' (e.g., Pickering and Garrod, 2004; Gann and Barr, 2014), they may decide to mention the color of an atypically colored object because this is salient to the addressees as well.

Although the face-to-face tasks in our experiments do not offer conclusive evidence in this discussion, there are reasons to believe that overspecified atypical color attributes are beneficial for addressees. For example, a visual world study by Huettig and Altmann (2011; Experiment 3) suggests that listeners tend to look for objects in typical colors when this color is not specified for them. When listeners hear a word that refers to an object with a prototypical color (even though this color is not mentioned), their visual attention shifts toward objects that have this particular color. So, listeners likely benefit from color being included in a referring expression when this color is not in line with their expectations about the object they search for. Similar suggestions come from work in visual search, which gives reasons to assume that listeners who are informed about specific details of the target, such as its color, find the target more efficiently in real-world scenes (e.g., Malcolm and Henderson, 2009, 2010).

The addressed literature is less clear on how the interaction with shape diagnosticity that we report in Experiment 2 might translate to effects for addressees. As shape diagnosticity moderates effects of color atypicality on reference production, one could speculate that a similar moderation applies to the addressees' task of identifying the intended target object. The object recognition literature suggests that color is relatively less instrumental in recognition for complex-shaped objects (e.g., Tanaka and Presnell, 1999; Bramão et al., 2011a), so for these objects listeners can rely more on shape-based cues in their visual search for the intended target object. Conversely, for simple-shaped objects color is a relatively more useful cue for finding these objects in a visual context (i.e., color is particularly instrumental to find the target in visual search). For example, when addressees search for a tomato, redness is a more relevant cue compared to when they search for a lobster. From this it follows (speculatively) that being informed about the color of the target object being atypical is more beneficial for listeners when they search for simply shaped objects, compared to when they search for objects for which shape is more instrumental for identifying the target. More research is needed to explore the effects of mentioning color on visual search, and interactions with color typicality and shape diagnosticity.

Acknowledgments

We thank many of our colleagues for their valuable comments on earlier versions of this work. We also thank Nanne van Noord for applying the computational physical salience estimations, and Eveline van Groesen for her assistance in Experiment 1. We are grateful to the editor and the reviewers for their valuable comments and suggestions.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00935

References

Andreopoulos, A., and Tsotsos, J. K. (2013). 50 Years of object recognition: directions forward. *Comput. Vis. Image Underst.* 117, 827–891. doi: 10.1016/j.cviu.2013.04.005

Arnold, J. E. (2008). Reference production: production-internal and addressee-oriented processes. *Lang. Cogn. Process.* 23, 495–527. doi: 10.1080/01690960801920099

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Becker, M. W., Pashler, H., and Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *J. Exp. Psychol. Hum. Percept. Perform.* 33, 20–30. doi: 10.1037/0096-1523.33.1.20

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147. doi: 10.1037/0033-295X.94.2.115

Bramão, I., Reis, A., Petersson, K. M., and Faísca, L. (2011a). The role of color information on object recognition: a review and meta-analysis. *Acta Psychol.* 138, 244–253. doi: 10.1016/j.actpsy.2011.06.010

Bramão, I., Inácio, F., Faísca, L., Reis, A., and Petersson, K. M. (2011b). The influence of color information on the recognition of color diagnostic and noncolor diagnostic objects. *J. Gen. Psychol.* 138, 49–65. doi: 10.1080/00221309.2010.533718

Brennan, S. E., and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. Exp. Psychol. Learn. Mem. Cogn.* 22, 1482–1493. doi: 10.1037/0278-7393.22.6.1482

Clark, H. H. (1996). *Using Language*. Cambrigde: Cambridge University Press. doi: 10.1017/CBO9780511620539

Clarke, A. D., Elsner, M., and Rohde, H. (2013). Where's Wally: the influence of visual salience on referring expression generation. *Front. Psychol.* 4:329. doi: 10.3389/fpsyg.2013.00329

Coco, M. I., and Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cogn. Sci.* 36, 1204–1223. doi: 10.1111/j.1551-6709.2012.01246.x

Conklin, E. J., and McDonald, D. D. (1982). "Salience: the key to the selection problem in natural language generation," in *Proceedings of the 20th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, 129–135. doi: 10.3115/981251.981287

Dale, R., and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902_3

Dale, R., and Viethen, J. (2009). "Referring expression generation through attribute-based heuristics," in *Proceedings of the 12th European Workshop* *on Natural Language Generation*, Athens, 58–65. doi: 10.3115/1610195.1610204

Engelhardt, P. E., Barış Demiral, Ş., and Ferreira, F. (2011). Over-specified referring expressions impair comprehension: an ERP study. *Brain Cogn.* 77, 304–314. doi: 10.1016/j.bandc.2011.07.004

Erdem, E., and Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *J. Vis.* 13, 11. doi: 10.1167/13.4.11

Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336, 998–998. doi: 10.1126/science.1218633

Fukumura, K., and van Gompel, R. P. (2012). Producing pronouns and definite noun phrases: do speakers use the addressee's discourse model? *Cogn. Sci.* 36, 1289–1311. doi: 10.1111/j.1551-6709.2012.01255.x

Fukumura, K., van Gompel, R. P. G., and Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *Q. J. Exp. Psychol.* 63, 1700–1715. doi: 10.1080/17470210903490969

Gann, T. M., and Barr, D. J. (2014). Speaking from experience: audience design as expert performance. *Lang. Cogn. Neurosci.* 29, 744–760. doi: 10.1080/01690965.2011.641388

Gauthier, I., and Tarr, M. J. (1997). Becoming a "greeble" expert: exploring mechanisms for face recognition. *Vision Res.* 37, 1673–1682. doi: 10.1016/S0042-6989(96)00286-6

Horton, W. S., and Gerrig, R. J. (2005). Conversational common ground and memory processes in language production. *Discourse Process.* 40, 1–35. doi: 10.1207/s15326950dp4001_1

Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117. doi: 10.1016/0010-0277(96)81418-1

Huettig, F., and Altmann, G. T. (2011). Looking at anything that is green when hearing "frog": how object surface colour and stored object colour knowledge influence language-mediated overt attention. *Q. J. Exp. Psychol.* 64, 122–145. doi: 10.1080/17470218.2010.481474

Humphreys, G. W., Riddoch, M. J., and Quinlan, P. T. (1988). Cascade processes in picture identification. *Cogn. Neuropsychol.* 5, 67–104. doi: 10.1080/02643298808252927

Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007

Keysar, B., Barr, D. J., and Horton, W. S. (1998). The egocentric basis of language use: insights from a processing approach. *Curr. Dir. Psychol. Sci.* 7, 46–50. doi: 10.1111/1467-8721.ep13175613

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Koolen, R., Goudbeek, M., and Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cogn. Sci.* 37, 395–411. doi: 10.1111/cogs.12019

Koolen, R., Houben, E., Huntjens, J., and Krahmer, E. (2014). "How perceived distractor distance influences reference production: effects of perceptual grouping in 2D and 3D scenes," in *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, Quebec.

Krahmer, E., and Van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI_a_00088

Kuhlen, A. K., and Brennan, S. E. (2013). Language in dialogue: when confederates might be hazardous to your data. *Psychon. Bull. Rev.* 20, 54–72. doi: 10.3758/s13423-012-0341-8

Landragin, F. (2004). Saillance physique et saillance cognitive. *Cogn. Représentation Langage* 2.

Malcolm, G. L., and Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: evidence from eye movements. *J. Vis.* 9, 8.1–8.13. doi: 10.1167/9.11.8

Malcolm, G. L., and Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *J. Vis.* 10, 4.1–4.11. doi: 10.1167/10.2.4

Mapelli, D., and Behrmann, M. (1997). The role of color in object recognition: evidence from visual agnosia. *Neurocase* 3, 237–247. doi: 10.1080/13554799708405007

McRae, K., Cree, G., Seidenberg, M., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* 37, 547–559. doi: 10.3758/BF03192726

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h0043158

Mitchell, M. (2013). *Generating Reference to Visible Objects*. Ph.D. thesis, University of Aberdeen, Aberdeen.

Mitchell, M., Reiter, E., and Deemter, K. V. (2013a). "Typicality and object reference," in *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (*CogSci*), Berlin.

Mitchell, M., Reiter, E., and Deemter, K. V. (2013b). "Attributes in visual object reference," in *Proceedings of Bridging the Gap between Cognitive and Computational Approaches to Reference* (*PRE-CogSci*), Berlin.

Naor-Raz, G., Tarr, M. J., and Kersten, D. (2003). Is color an intrinsic property of object representation? *Perception* 32, 667–680. doi: 10.1068/p5050

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 89–110. doi: 10.1515/ling.1989.27.1.89

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–190. doi: 10.1017/S0140525X04000056

Price, C. J., and Humphreys, G. W. (1989). The effects of surface detail on object categorization and naming. *Q. J. Exp. Psychol. A* 41, 797–827. doi: 10.1080/14640748908402394

Rosch, E. (1975). Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.* 104, 192–233. doi: 10.1037/0096-3445.104.3.192

Rosch, E., and Mervis, C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cogn. Psychol.* 7, 573–605. doi: 10.1016/0010-0285(75)90024-9

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *J. Psycholinguist. Res.* 32, 3–23. doi: 10.1023/A:1021928914454

Tanaka, J., and Presnell, L. (1999). Color diagnosticity in object recognition. *Percept. Psychophys.* 2, 1140–1153. doi: 10.3758/BF03207619

Tanaka, J., Weiskopf, D., and Williams, P. (2001). The role of color in high-level vision. *Trends Cogn. Sci.* 5, 211–215. doi: 10.1016/S1364-6613(00)01626-0

Therriault, D., Yaxley, R., and Zwaan, R. (2009). The role of color diagnosticity in object recognition and representation. *Cogn. Process.* 10, 335–342. doi: 10.1007/s10339-009-0260-4

Van Deemter, K., Gatt, A., Sluis, I. V. D., and Power, R. (2012a). Generation of referring expressions: assessing the incremental algorithm. *Cogn. Sci.* 36, 799–836. doi: 10.1111/j.1551-6709.2011.01205.x

Van Deemter, K., Gatt, A., Van Gompel, R. P. G., and Krahmer, E. (2012b). Toward a computational psycholinguistics of reference production. *Top. Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01187.x

Viethen, J., Dale, R., and Guhe, M. (2014). Referring in dialogue: alignment or construction? *Lang. Cogn. Neurosci.* 29, 950–974. doi: 10.1080/01690965.2013.827224

Viethen, J., Goudbeek, M., and Krahmer, E. (2012). "The impact of colour difference and colour codability on reference production," in *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (*CogSci*), Sapporo.

Wardlow Lane, L., Groisman, M., and Ferreira, V. S. (2006). Don't talk about pink elephants! Speakers' control over leaking private information during language production. *Psychol. Sci.* 17, 273–277. doi: 10.1111/j.1467-9280.2006.01697.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 February 2016
doi: 10.3389/fpsyg.2016.00103

Talking about Relations: Factors Influencing the Production of Relational Descriptions

*Adriana Baltaretu *, Emiel J. Krahmer, Carel van Wijk and Alfons Maes*

Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, Netherlands

In a production experiment (Experiment 1) and an acceptability rating one (Experiment 2), we assessed two factors, spatial position and salience, which may influence the production of relational descriptions (such as "the ball between the man and the drawer"). In Experiment 1, speakers were asked to refer unambiguously to a target object (a ball). In Experiment 1a, we addressed the role of spatial position, more specifically if speakers mention the entity positioned leftmost in the scene as (first) relatum. The results showed a small preference to start with the left entity, which leaves room for other factors that could influence spatial reference. Thus, in the following studies, we varied salience systematically, by making one of the relatum candidates animate (Experiment 1b), and by adding attention capture cues, first subliminally by priming one relatum candidate with a flash (Experiment 1c), then explicitly by using salient colors for objects (Experiment 1d). Results indicate that spatial position played a dominant role. Entities on the left were mentioned more often as (first) relatum than those on the right (Experiments 1a–d). Animacy affected reference production in one out of three studies (in Experiment 1d). When salience was manipulated by priming visual attention or by using salient colors, there were no significant effects (Experiments 1c, d). In the acceptability rating study (Experiment 2), participants expressed their preference for specific relata, by ranking descriptions on the basis of how good they thought the descriptions fitted the scene. Results show that participants preferred most the description that had an animate entity as the first mentioned relatum. The relevance of these results for models of reference production is discussed.

Keywords: reference production, relatum, spatial position, animacy, perceptual salience, attention capture, relational descriptions, referring expressions

Edited by:
Albert Gatt,
University of Malta, Malta

Reviewed by:
Thora Tenbrink,
Bangor University, UK
Alasdair Daniel Francis Clarke,
University of Aberdeen, UK

****Correspondence:***
Adriana Baltaretu
a.a.baltaretu@tilburguniversity.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *14 July 2015*
Accepted: *19 January 2016*
Published: *09 February 2016*

Citation:
Baltaretu A, Krahmer EJ, van Wijk C and Maes A (2016) Talking about Relations: Factors Influencing the Production of Relational Descriptions. Front. Psychol. 7:103. doi: 10.3389/fpsyg.2016.00103

1. INTRODUCTION

Human speakers have a rich repertoire for referring to objects in visual scenes. For example, if you want to buy a ball from the toy store, the shop assistant could help you find it among other balls by referring to intrinsic attributes (e.g., color, *the red ball*) or extrinsic ones (e.g., location, *the ball between the doll and the train*). An object's location can be described in relation to one's body and to other objects or to environmental features (Levinson, 1996). In the current work, we focus on referential choices when describing external relations (Levinson, 2003; Tenbrink, 2011) where an object is the target, while other object(s) serve as the relatum. The target is sometimes referred to as the locatum, figure or located object, whereas the relatum is also known as ground, reference location or landmark. In the previous example, the ball represents the target and it is described in relation to two relata objects, the doll and the train.

Compared to intrinsic attributes (such as color), there are few studies in the referring expressions generation field analyzing how extrinsic attributes (such as location) are used in order to refer unambiguously to a target object (for a review, see Krahmer and Van Deemter, 2012). When talking about location, speakers describe where the target object is positioned in space. Far from being a trivial feature, space is a pervasive dimension in language and cognition. For example, we map time onto space (e.g., Boroditsky, 2000), make use of space in gestures (e.g., Gentner et al., 2013), in discourse (e.g., Lakoff and Johnson, 2008), and in actions (e.g., Kirsh, 1995). Crucially, humans employ location in a meaningful way in different forms of descriptions and visualizations. It is natural to refer to an object's location in a variety of situations, thus anchoring the conversation topic in the spatio-temporal context (Levelt, 1993, p.51). Such situations are, among other things, route direction production, interaction with conversational agents, visual communication (e.g., maps and graphs) within various disciplines (e.g., architecture, geosciences, engineering, etc., for a review, see Tversky, 2011).

Pervasive use of spatial relations in real life communication makes it necessary to develop referring expression generation algorithms that can handle such reference. These algorithms (e.g., the Incremental Algorithm, Dale and Reiter, 1995; the Graph-Based Algorithm, Krahmer et al., 2003) have a key role in natural language generation, enabling machines to make informed choices and to refer to objects in a more human-like manner (van Deemter et al., 2012; Gatt et al., 2014; Dos Santos Silva and Paraboni, 2015). Though we know little of the situations when relational descriptions are spontaneously produced and preferred over intrinsic attributes, there are communicative contexts in which relations are an efficient and relevant strategy [like in route directions or in scenes with many (similar) objects]. Recent studies have shown that speakers often produce relational descriptions in order to single target objects out of other objects in a visual scene (Clarke et al., 2013a; Kazemzadeh et al., 2014). When both intrinsic and extrinsic attributes are available, people tend to mention location even when this attribute is not necessary for producing a unique object description (Viethen and Dale, 2008). Listeners seem to benefit from this type of reference as well (Arts et al., 2011; Paraboni and van Deemter, 2014). Currently, spatial relations represent a major challenge for referring expressions generation algorithms, as we know little about the situations in which speakers employ them in the context of identification. To further develop these algorithms, more input from studies on human reference is needed.

In this series of studies, we focus on human reference production in spatial relational descriptions. In visual scenes, several entities can be in the proximity of the target and each one of them could be a potential relatum. In our previous example, the shop assistant could either refer to the target as, for example, *the ball in front of the doll* (using a single relatum) or *the ball between the doll and the train* (using two relata). In the first description, which we call *the single-relatum formulation*, the question is what causes speakers to mention one of the objects. In the second strategy, *the two-relata formulation*, we question what causes speakers to mention one of the objects as first relatum. In the two-relata formulation, we consider important the order in which entities are mentioned. Word order choices have been previously suggested to reflect speaker's referential preferences (Goudbeek and Krahmer, 2012) and the ease with which these entities are processed (Bresnan et al., 2007; Onishi et al., 2008; Jaeger and Tily, 2011).

While the study of spatial relations in the field of referring expression generation is a topic largely unexplored, in the field of spatial cognition there have been numerous studies concerned with principles that govern relatum object selection (e.g., Barclay and Galton, 2008; Miller et al., 2011; Barclay and Galton, 2013), the choice of adequate spatial prepositions based on geometric and functional characteristics of the objects (e.g., Carlson-Radvansky et al., 1999; Coventry and Garrod, 2004) and the influence of frames of reference on relatum selection (e.g., Carlson-Radvansky and Radvansky, 1996; Levinson, 2003; Taylor and Rapp, 2004; Tenbrink, 2007). Various factors might affect the selection of a relatum object. Compared to target objects, relata are described as larger, closer to the target, geometrically more complex (Barclay and Galton, 2013) as well as more familiar, expected, more immediately perceivable (Talmy, 2003).

In this series of studies, we seek to investigate speakers' referential choices, aiming thereby to provide further insight for REG algorithms. Most studies mentioned above focus on the problem of localization, as opposed to identification (Tenbrink, 2005; Dos Santos Silva and Paraboni, 2015). In localization tasks speakers are restricted to refer to already agreed upon objects (e.g., the target and relatum are given and a priori labeled as, for example "cup"), based solely on their spatial locations. On the other hand, freely producing a referring expression (like "the cup between the plate and the kettle") is a matter of choosing attributes of the target (including its spatial position), to help the addressee identify a target object out of several candidates. Comparisons between identification and localization tasks have been previously addressed (Tenbrink, 2005; Moratz and Tenbrink, 2006; Vorwerg and Tenbrink, 2007). In general, descriptions seem to be more detailed when the target needs to be localized, rather than identified. Factors to influence reference production (e.g., spatial biases, conceptual and visual salience) have been addressed to a lesser extent.

It is generally assumed that if an object is salient, it can grab visual attention, and thus is likely to be selected and mentioned as relatum (Beun and Cremers, 1998; Tversky et al., 1999). A number of visual factors have been identified as important cues for salience, such as size, color, orientation, foregrounding, animacy (for a review, see Wolfe, 1994; Parkhurst et al., 2002; Kelleher et al., 2005; Coco and Keller, 2015), but little is known about how these and other cues influence reference production. The goal of the current research is to examine two factors previously shown to influence language production and comprehension in general, yet understudied in reference production: spatial position and salience.

1.1. Spatial Position: A Left-to-Right Preference?

Referring to a relatum may be influenced by a factor present in any visual scene: the position of the object in the scene. Different types of evidence suggest there might be a bias to choose objects placed in specific locations. Speakers choose and mention spatially aligned and proximate objects as relata (e.g., Craton et al., 1990; Hund and Plumert, 2007; Viethen and Dale, 2010; Miller et al., 2011). Yet, when several objects are in the vicinity of the target, all similarly aligned, would spatial features continue to influence reference production? We assume that it does, and objects on the left of the target would be mentioned more often as relatum than objects on the right. This prediction is based on findings from various disciplines as follows.

The speaker's attention might be guided by different factors toward specific regions of the scenes. One line of research suggests that oculomotor biases (the amplitude and direction of saccades—movements of the eye between fixation points) are an important predictor for the location where speakers initially direct their attention (e.g., Tatler and Vincent, 2009; Kollmorgen et al., 2010). One well known, image independent bias is the tendency to look at the center of visual stimuli during image exploration (for a review, see Clarke et al., 2013a). Besides this bias, there is also evidence for a horizontal spatial bias (sometimes referred to as "pseudoneglect"). People initially execute more often leftward than rightward saccades, irrespective of the content of the image, across different tasks (free viewing, memorization, scene search, Foulsham et al., 2013; Ossandón et al., 2014). This asymmetry seems to affect memory, with left positioned objects being better remembered than right positioned ones (Dickinson and Intraub, 2009).

Converging evidence comes from cross-cultural psychology research where the left-to-right bias is considered to be a result of the scanning routines employed during reading and writing. The directionality of the language system has an impact on visual attention, memory, and spatial organization (Chan and Bergen, 2005). For instance, when participants with a left-to-right language system (in this case: French) were asked to mark the middle of a straight line, they usually misplaced the mark to the left of the objective middle, while participants with a right-to-left language system (Hebrew) misplaced the mark to the right (Chokron and Imbert, 1993). Such a bias is shown from a young age in graphical representations of spatial and temporal relations (Tversky et al., 1991). This implies that, at least in western cultures, people "read" visual scenes from left to right and that the left-to-right bias might be a habit acquired by systematically using a language system.

The directionality of the writing system seems to affect cognitive linguistic processes. In picture description tasks, speakers of left-to-right languages tend to scan, describe and remember items from left to right (Taylor and Tversky, 1992; Meyer et al., 1998). Speakers of different writing systems show different patterns of sentence production. For example, in a sentence-picture matching task, speakers of a language with a left-to-right (in this case: Italian) system tended to choose visual scenes with the agent placed on the left of the patient, those of a language with a right-to-left system (Arabic) preferred scenes with the agent placed on the right of the patient (Maass and Russo, 2003; Chan and Bergen, 2005). Not only the writing system, but also the dominant frame of reference of the language, might affect the order in which speakers refer to entities in visual scene. For example, when using a relative frame of reference, to perceive that something is "on the left," the speaker would project his viewpoint onto the scene (Levinson, 2003). Bilingual speakers of Spanish (relative frame of reference) and Yucatec (no dominant frame of reference), show a bias to start with the left object in the scene when using Spanish, but not when doing this task in Yucatec (Butler et al., 2014).

The left-to-right bias was also observed in clinical populations. Participants suffering from agrammatism, an aphasic syndrome, presented a similar left-to-right bias both in language production (describing visual scenes) and comprehension (matching sentences with pictures, Chatterjee, 2001). In addition, studies in the psychology of art suggest that reading habits influence visual preferences: participants preferred pictures possessing the same directionality as their reading system (Chokron and De Agostini, 2000).

Given the evidence for a left-to-right bias, there might be a tendency for speakers to mention relata based on their position in the scene. For example, in **Figure 1**, speakers could refer to the target as in (a) *the ball in front of the bookshelf*, (b) *the ball in front of the clock* or (c) *the ball between the bookshelf and the clock*. These three descriptions were considered valid for identification and classified in two formulation preferences: the single-relatum formulation (descriptions a and b) and the two-relata formulation (description c). When only one object was mentioned, we considered it to reflect the speakers' preference for a relatum candidate. In case both entities were mentioned, we took into account the order of mentioning. If a left-to-right bias plays a role in reference production, we expect entities left of the target to be mentioned more often as relatum (as in *a*) or mentioned more often as the first relatum (as in *c*). However, a spatial bias, might not be the sole factor that influences relatum reference. In the following section, we review evidence for other factors that potentially contribute to the salience of relatum candidates.

1.2. Salience

Salience is generally considered an important factor for reference production. The objects' salience captures visual attention and entities in focus of attention during utterance planning have

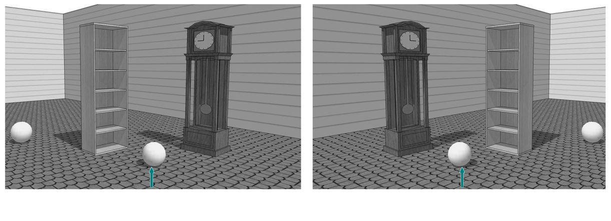


FIGURE 1 | Experimental stimulus with inanimated object (bookshelf) on the left (A) and the right (B) of the target.

higher chances of being mentioned (Beun and Cremers, 1998; Gleitman et al., 2007). In the present study, salience (the property of being noticeable or important) is operationalized in two ways.

We distinguish between conceptual and visual salience. By conceptual salience, we refer to the ease of activation of mental representations caused by knowledge-based conceptual information (or "accessibility" in Bock and Warren, 1985; Ariel, 1990). There are several properties of the referent that contribute to its conceptual salience (e.g., linguistic properties, such as the syntactic position a referent occupies; context, such as the preceding discourse; intrinsic properties, such as animacy, etc.). In this study we focus on animacy: whether an entity is conceptualized as living or not (Vogels et al., 2013; Coco and Keller, 2015). In contrast, by visual salience we touch on two different aspects: perceptual salience and visual priming. By perceptual salience, we refer to bottom-up, stimulus-driven signals that attract visual attention to areas of the scene that are sufficiently different from the surroundings (Itti and Koch, 2001). For example, a perceptually salient object is an object that has a unique color compared to the rest of the scene. Moreover, entities can become salient when visual attention is guided toward them, for example by using attention priming techniques (Gleitman et al., 2007). Below we discuss these types of salience in more detail.

1.2.1. Conceptual Salience

Animacy is a basic conceptual feature of objects and there are reasons to believe that it may affect the production of relational descriptions. First, animacy has been shown to influence the allocation of visual attention. Humans prioritize the visual processing of animate objects over inanimate ones (Kirchner and Thorpe, 2006; New et al., 2007; Fletcher-Watson et al., 2008). Both visual representations of the face and the human body have the ability to capture the focus of attention, even when attention is occupied by another task (Downing et al., 2004). Compared to inanimate objects, animate entities are more likely to be fixated and named (Clarke et al., 2013b; for a review, see Henderson and Ferreira, 2013).

Second, animacy is known to play a key role in reference production (Clark and Begun, 1971; McDonald et al., 1993). Animate entities are conceptually highly accessible, thus, retrieved and processed more easily than inanimate entities (Prat-Sala and Branigan, 2000). This can influence word ordering, as there is a strong tendency for the animate entities to occupy more prominent syntactic positions (e.g., in the beginning of a structure) and grammatical functions (e.g., subject role) (e.g., Bock et al., 1992; McDonald et al., 1993; Prat-Sala and Branigan, 2000; Branigan et al., 2008). Additionally, compared to inanimate referents, animates are mentioned more frequently and are more likely to be pronominalized (e.g., Fukumura and van Gompel, 2011).

Given that utterance planning is influenced by conceptual factors and that animacy has a privileged role in language production, we could expect animate entities to be mentioned as relatum (or as first relatum) more often than inanimate ones due to their conceptual salience, irrespective of their position with respect to the target. In general, there is little evidence that animacy could influence relatum choice. The few studies that looked at this, directly or indirectly, do not present a consistent picture. Under specific circumstances, de Vega et al. (2002) report that relata can be animate, but only when included in a construction using the preposition *behind* [the animate entity]. Congruent evidence was found in a large English corpus of referring expressions elicited with complex naturalistic scenes. Speakers were shown an image with an outlined object and provided with a text box in which to write a referring expression. When speakers decided to produce spatial relational descriptions, the most frequent relata objects were people and some entities positioned in the background, such as trees and walls (Kazemzadeh et al., 2014). Taylor et al. (2000), however, argue that animate entities should be disfavored as relata due to their mobility.

1.2.2. Visual Salience

Reference production was shown to be sensitive to both visual priming (e.g., a short flash at the target location, Gleitman et al., 2007) and perceptual salience cues, such as uniquely colored objects (Pechmann, 1989; Belke and Meyer, 2002).

Priming participants' initial gaze to a specific area of a scene has been claimed to influence grammatical role assignment and word order (Gleitman et al., 2007). When visual attention is guided toward it, an object is more likely to be mentioned in the beginning of a description or relation (in a prominent grammatical role, such as subject, or in a prominent position in the utterance). As far as we know, no studies looked into effects of attention manipulation on spatial relational descriptions. Reference production can be influenced by very basic, implicit attention-grabbing cues. Gleitman et al. (2007) report that presenting a flash shortly before displaying a scene, systematically redirected the gaze of the participants to the location of a specific object (occurring at the location of the flash), which later received a privileged position in the sentence structure. The short duration of the flash ensured that participants remained unaware of the manipulation, while their gaze was attracted to the cued location in an implicit manner.

A similar approach has been used for the study of spatial relational descriptions (*X is left of Y*). Forrest (1996) drew speakers' attention to the location of an object, prior to the scene presentation. Unlike Gleitman et al. (2007), she used an explicit visual cue, a flash that lasted long enough to be noticed by the participants. This explicit visual cue influenced speakers' description as well: the object which appeared in the primed location generally received a more prominent place in the beginning of the sentence.

Apart from priming, properties of the stimulus may play a crucial role in guiding the eyes. Perceptual salience is a factor known to influence visual attention (for review, see Tatler et al., 2011) and reference production (Myachykov et al., 2011; Clarke et al., 2013b; Coco and Keller, 2015). Perceptual salience is a characteristic of parts of a scene (objects or regions), that appear to stand out relative to their neighboring parts and there are several models to account for this phenomenon (for a review, see Borji and Itti, 2013). Most models use image features, such as color, contrast, orientation and motion and make center-surround operations to compare the statistics of image features at a given location to the statistics in the surrounding area (Borji and Itti, 2013).

Among these features, color has been shown to capture visual attention (Folk et al., 1994; Parkhurst et al., 2002), irrespective of the observers' task (Theeuwes, 1994). In general, color enhances object recognition (for a review, see Tanaka et al., 2001) and uniquely colored items are detected faster than other objects in the scene, regardless of the amount of distractors (Treisman and Gelade, 1980; D'Zmura, 1991).

In general, scholars suggest that explicit perceptual features (such as color, size, shape) may contribute to relatum selection (e.g., Barclay and Galton, 2008), yet there are almost no experimental studies which try to disentangle the effects of these features. Regarding the influence of color on relatum selection and reference, prior results are equivocal (Miller et al., 2011, Viethen et al., 2011). Yet, in reference production studies, color is probably the attribute mentioned most frequently. In reference tasks, color is considered to have a high pragmatic value (Belke and Meyer, 2002; Davies and Katsos, 2009). Speakers mention it even when this information is not needed for identification (Koolen et al., 2011; Westerbeek et al., 2015). In complex scenes, reference to both target and relatum objects is affected by perceptual salience (a composite measure of color and other low level visual features), visual complexity (clutter), size and proximity (Clarke et al., 2013a). Clarke et al. (2013a) note that relatum objects were chosen based on their size and saliency; while references to less salient target objects included a higher number of relata.

Moreover, the order in which objects are mentioned in a relational description may be sensitive to perceptual salience as well. In visual domains, speakers can mention target and relatum objects in different orders. Elsner et al. (2014) report that speakers employed complex word orders such as starting with (a) the target, (b) the relatum or by giving information about the target in multiple phrases intertwined with relatum references. For example, if the target was a person (target in **bold**, relatum in *italics*), speakers could say (a) **man** closest to *the rear tyre of the van*, (b) near *the hut that is burning*, there is **a man holding a lit torch in one hand, and a sword in the other** or (c) there is **a person standing** in *the water* **wearing a blue shirt and yellow hat** (Elsner et al., 2014, p. 522). These relations were more likely to start with the perceptually salient object.

Given these findings, we could expect objects to be mentioned as (first) relatum if they are placed in a cued location or if they are perceptually salient.

1.3. The Current Studies

Spatial position (left-to-right bias), conceptual salience (animacy), and visual salience (attention capture cues or scene based perceptual cues) all influence what is being looked at (Kollmorgen et al., 2010) and possibly mentioned (Coco and Keller, 2015). We study if and to what extent these factors influence referential choices in spatial relational descriptions.

This paper presents two experiments consisting of several parts that test the influence of these factors on relatum reference in an identification task. In Experiment 1a, we started by determining if there was a spatial bias when mentioning a relatum. We start with a basic language elicitation task that did not include any experimental factors. Its purpose was to check for a left bias in reference production. In this language elicitation task, we manipulated the position of two inanimate relatum candidates. Entities placed on the left of the target were expected to be mentioned as (first) relatum more often than those placed on the right. We took spatial position as a baseline and continued investigating the effect of salience on referential choices. Conceptual salience was manipulated by adding one animate entity in each scene (Experiment 1b). Animate entities were expected to be preferred as relatum. Visual salience was manipulated by priming attention toward a relatum candidate with a short flash (Experiment 1c) or explicitly with a unique color (Experiment 1d). Salient entities were expected to be preferred as relatum. Additionally, the listeners' preference for relata was tested, by asking participants to rank relational descriptions starting with the one that, according to them, "best fits" the scene (Experiment 2). Descriptions that have an animate entity as (first) relatum were expected to be ranked higher.

We explored these predictions across a production experiment (four parts) and in an acceptability rating experiment, and in doing so some factors may be included in several parts of these experiments (for example, the effect of spatial position is analyzed in Experiments 1, 2, animacy in Experiments 1b–d and in Experiment 2, visual salience in Experiments 1c–d). Whether speakers mentioned the left entity as (the first) relatum was tested by comparing the chance of naming the left item with random chance (0.50) using an one-sample t-test and possible interactions between the experimental factors were evaluated using analysis of variance (ANOVA) tests[1].

Finally, the current studies were carried out in accordance with the recommendations of APA guidelines for conducting experiments, the Netherlands Code of Conduct for Scientific Practice and the Code for Use of Personal Data in Scientific Research (KNAW). The studies were approved by the ethics committee at Tilburg University and all participants gave written consent to the use of their data.

2. EXPERIMENT 1—REFERENCE PRODUCTION

2.1. Experiment 1a—Position

2.1.1. Participants

Thirty native Dutch undergraduates from Tilburg University participated in this study for partial course credits. Data from four speakers were discarded on the basis of task misunderstanding. The final sample consisted of 26 participants (11 female, mean age 20.19).

[1] The Huynh-Feldt epsilon value was pretty close to 1 in all the analyses, indicating that there was no need for adjustments of the degrees of freedom.

2.1.2. Materials

The stimuli consisted of 48 grayscale scenes (12 experimental stimuli). The experimental stimuli scenes included a target item marked with an arrow (a ball), a distractor object (a ball identical to the target) in order to prevent an easy identification strategy using type only, and two relatum candidates (both inanimates). These items were eight everyday objects (such as wardrobes), easily identifiable, with a clear front/back axis and of roughly equal size, randomly coupled in pairs (see **Figure 1**). Filler stimuli were used to have a larger visual diversity (they included both inanimate and animate objects) and to allow participants to use a wider range of identification strategies (type, location and size). All the objects (8 animate and 8 inanimate) were pretested with a group of 10 participants, who were presented with pictures similar to the ones used in this study. They had to name the inanimate objects, as well as the gender and profession of animate objects. An inanimate object was included in the experimental stimuli if (1) it was referred to with the same noun in a minimum of 50% of the cases, and (2) if the other nouns used to refer to it, were compound nouns such as in "kast"–"ladenkast" (drawer). An animate object was chosen if (1) the character's gender was recognized in all cases and (2) if the character's profession was recognized in 80% of the cases. The scenes were created using Google SketchUp 8 (3D Warehouse library).

2.1.3. Procedure

Participants were instructed to verbally refer to an object marked with an arrow in such a way that the next participant (a fictitious listener) could draw the arrows on a new set of identical pictures (language: Dutch). The goal of this instruction was to avoid participants to produce ambiguous references (for a similar procedure see Koolen et al., 2011; Clarke et al., 2013a). Participants saw each entity in three different pictures, paired every time with a different object. The materials were divided across two presentation lists, so that each participant would see each object combination only once. The position of each object and the position of the distractor ball were individually counterbalanced (half of the times they appeared on the left of the scene and half of the times on the right of the scene). Descriptions such as *the ball in front of me* or *the ball on the left* were discouraged, by telling the speaker that the listener would receive the same image, but that it might be in a mirror version. The picture remained on the screen until the participant produced a description and pressed a button to continue. Each experimental trial was followed by 3 filler trials to prevent a carry-over effect. The study started with 3 practice trials followed by 48 experimental trials and lasted approximately 10 min.

2.1.4. Results and Discussion

We collected 312 descriptions (26 participants * 12 experimental stimuli). Participants were found to use one of two possible formulations: either mentioning a single relatum (e.g., *the ball in front of the bookshelf*) or both (e.g., *the ball in between the bookshelf and the clock*). In all the studies of Experiment 1, the participants were grouped based on their preference for the single-relatum or the two-relata formulation strategy. Some participants systematically used a single formulation strategy, while others used both. The grouping threshold was set by inspecting the distribution of the two-relata formulation in Experiment 1. The distribution appeared to be bimodal: one group had a score of maximum 100% (down to 80); the other group had a score of maximum 40% (down to 0). Every participant with a score of 80 or more was considered to opt for a two-relata formulation and all the other for a single-relatum formulation.

In Experiment 1a participants were found to use a single-relatum formulation ($N = 1$ participant, not analyzed further due to small sample size) or a two-relata formulation ($N = 25$ participants). Whether speakers mentioned the left entity as the first relatum was tested by comparing the chance of naming the left item with random chance (0.50) using an one-sample t-test. Speakers mentioned the left entity as first relatum 59% of the time (95% CI [0.525; 0.659], $SD = 0.16$). This result was statistically significant [$t_{(24)} = 2.857, p = 0.009; d = 0.57$].

The results showed a left bias in reference production, however there was only a small preference in starting with the left entity. This leaves room for other factors that could influence reference. Thus, in Experiments 1b–d, we added three experimental factors that contribute to the entity's salience, making the entities "stand out" in the scene.

2.2. Experiment 1b—Conceptual Salience: Animacy

2.2.1. Participants

Fifty three native Dutch undergraduates from Tilburg University participated in this study as speakers for partial course credits. Due to technical problems, speech data of four participants were not analyzed; the final sample included 49 participants (11 males, mean age 21.2 years).

2.2.2. Materials

The stimuli consisted of 96 grayscale scenes (24 experimental stimuli). For these scenes, we used the same animate and inanimate objects described in Experiment 1a. The experimental stimuli consisted of a target and a distractor ball and two relatum candidates, one animate and one inanimate object of roughly equal size (see **Figure 2**). From 64 possible animate–inanimate combinations, 24 couples were randomly chosen. Filler stimuli were similar to the ones used in Experiment 1a.

2.2.3. Procedure

As in Experiment 1a.

2.2.4. Results and Discussion

Speakers produced 1176 descriptions (49 participants * 24 experimental stimuli). Participants were found to use one of two possible formulations: either mentioning a single relatum ($N = 12$) or both relata ($N = 37$). Whether speakers mentioned the left entity as the first relatum was tested by comparing the chance of naming the left item with random chance (0.50) using an one-sample t-test. The chance of mentioning the left entity as first relatum was 59% [two-sided 95% CI [0.55, 0.64], $SD = 0.17$, $t_{(47)} = 3.91, p < 0.001, d = 0.75$].

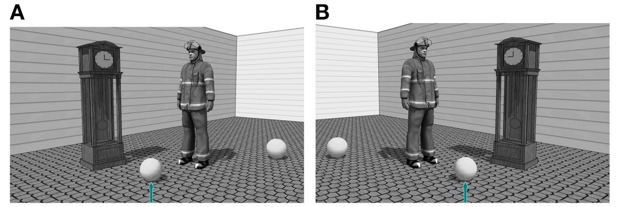


FIGURE 2 | Experimental stimulus with animated object (firefighter) on the right (A) and the left (B) of the target object.

Whether animacy overruled the left bias was tested with an ANOVA test, having Position of the Animate in the scene (2 levels: animate left, animate right) as a within subjects factor, and Participant Formulation Preference (2 levels: single-relatum, two-relata) as a between subjects factor. The ANOVA test revealed no statistically significant effect of Position of the Animate ($F < 1$) or of Participant Formulation Preference ($F < 1$) and no interaction between these factors ($F < 1$).

These results suggest that animacy did not influence descriptions. The responses were not affected by word frequency: 90% of the participants referred to the animate entity using highly frequent words such as *de vrouw / de man* (the woman / the man). However, the position of the entity was found to affect reference to a greater extent, with left entities being more likely to be mentioned as (first) relatum than right ones. In Experiment 1c, we test the strength of this preference by manipulating the objects' visual salience.

2.3. Experiment 1c—Perceptual Salience: Flash

2.3.1. Participants

Thirty nine native Dutch undergraduates from Tilburg University participated in this study for partial course credits. Data from 27 participants (18 women, mean age 20.3 years) were used, the rest being discarded on the basis of having noticed the cue (1 participant), task misunderstanding (2 participants) or not using a relatum at all as in *the ball in the center* (9 participants).

2.3.2. Materials

Stimuli from Experiment 1b were used, slightly cropped so that the target object was placed exactly in the middle of the scene. The attention capture manipulation consisted of a black square, with an area of 0.5×0.5 degrees of visual angle, set against a white background (Gleitman et al., 2007).

2.3.3. Procedure

The procedure was identical to the one presented in Experiment 1a. In addition, an implicit visual attention cue was added. Participants sat approximately 60 cm from the monitor, set to 1680 × 1050 pixels, 60 Hz refresh rate. Before each trial, participants were first presented with a fixation cross on a white background (500 ms). The fixation cross was followed by the attention capture manipulation, which was presented for 65 ms, followed immediately by a stimulus scene. The position on screen of the attention-capture cue varied (in half of the trials the cue was positioned left and in half right).

2.3.4. Results and Discussion

Participants used one of the two formulations (single-relatum $N = 6$, two-relata $N = 21$). Whether spatial position influenced reference production was tested by comparing the chance of mentioning the left entity as first relatum with random chance, using one–sample t-test. The chance of mentioning the left entity as first relatum was 67% [two-sided 95% CI [0.59, 0.75], $SD = 0.19$, $t_{(26)} = 4.61$, $p < 0.001$, $d = 0.67$].

Whether animacy or attention priming overruled the left bias was analyzed with an ANOVA test, having the Position of the Animate (2 levels: animate left, animate right) and the Position of the Flash (2 levels: flash left, flash right) as within subjects factors, and Participant Formulation Preference (2 levels: single-relatum, two-relata) as a between subjects factor. The ANOVA test revealed no statistically significant main effects of the Position of the Animate ($F < 1$) or of the Position of the Flash ($F < 1$).

There was a main effect of Participant Formulation Preference [$F_{(1,\ 25)} = 6.66$, $p = 0.016$, $\eta_p^2 = 0.21$]. In the two-relata formulation, participants mentioned more often the left entity as (first) relatum ($M = 0.72$), than in the single-relatum formulation ($M = 0.51$). There were no significant interactions between these factors ($F < 1$).

Experiment 1c confirmed the speaker's preference to mention left entities first. There were no effects of the Position of the Animate or of the Position of the Flash. In Experiment 1d, we continue testing the strength of the left bias by making one of the entities perceptually salient.

2.4. Experiment 1d—Perceptual Salience: Color

2.4.1. Participants

Fifty five native Dutch undergraduates from Tilburg University participated in this study for partial course credits (32 women, mean age 22 years). One participant was discarded for never mentioning a relatum.

2.4.2. Materials

Stimuli from Experiment 1b were used. In addition, one relatum candidate in each picture had a unique color (red, blue, green or yellow), while all the other were grayscale (see **Figure 3**).

2.4.3. Procedure

As in Experiment 1a. The position of the colored relatum candidate and the position of the relatum candidates was counterbalanced across presentation lists.

2.4.4. Results and Discussion

Participants used one of the two possible formulations (43 participants mentioned both relata, 4 participants mentioned a single relatum) or produced mixed descriptions across trials with both single-relatum and two-relata formulations (7 participants). Due to small sample sizes, participants that opted for a single-relatum were grouped with those who used a mixed formulation and analyzed as a mixed formulation group.

Whether spatial position influenced reference production was tested by comparing the chance of mentioning the left item as

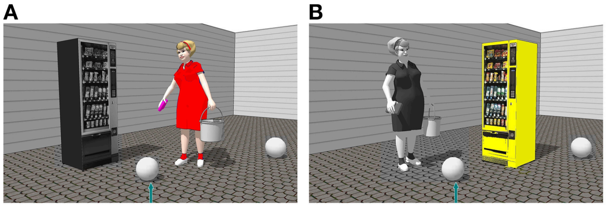


FIGURE 3 | Experimental stimulus with on the right of the target object in color (red) the animate object (A) and in color (yellow) inanimate object (B).

first relatum with random chance, using one–sample t-test. The chance of mentioning the left entity as first relatum was 61% [two-sided 95% CI [0.55, 0.66], $SD = 0.20$, $t_{(53)} = 3.81$, $p < 0.001$, $d = 0.47$].

Whether animacy or perceptual salience overruled the left bias was analyzed with an ANOVA test, having the Position of the Animate (2 levels: animate left, animate right) and the Position of the Colored entity (2 levels: colored left, colored right) as within subjects factors, and Participant Formulation Preference (2 levels: two-relata, mixed) as a between subjects factor.

There was no statistically significant effect of the Position of the Colored entity ($F < 1$).

There was a main effect of the Position of the Animate [$F_{(1,52)} = 18.645$, $p = 0.001$, $\eta_p^2 = 0.264$]. Participants mentioned the left entity as relatum more often when the animate entity was placed on the right of the scene ($M = 0.67$) than when the animate was placed on the left ($M = 0.43$).

There was a main effect of Participant Formulation Preference [$F_{(1,52)} = 6.613$, $p = 0.01$, $\eta_p^2 = 0.113$]. Participants mentioned the left entity as first relatum more often within a two-relata formulation ($M = 0.63$), than within a mixed one ($M = 0.47$).

There was an interaction between the Position of the Animate and Participant Formulation Preference [$F_{(1,\ 52)} = 4.183$, $p < 0.05$, $\eta_p^2 = 0.074$]. Speakers that used a two-relata formulation, mentioned the left entity as first relatum more often when the animate was on the right ($M = 0.70$) than on the left ($M = 0.57$). The same pattern of results was observed for speakers that used a mixed formulation (animate right $M = 0.65$, animate left $M = 0.29$). A split analysis showed that the general behavior of the two formulation groups is essentially the same, but the effect size is higher for the mixed formulation [$F_{(1,\ 10)} = 7.101$, $p = 0.024$, $\eta_p^2 = 0.415$], than for the two-relata one [$F_{(1,\ 42)} = 7.809$, $p = 0.008$, $\eta_p^2 = 0.157$].

Experiment 1d revealed that perceptual salience, namely entities with unique colors, did not influence reference production, while conceptual salience had a small influence.

Experiment 1 has examined the extent to which the production of spatial relational descriptions is influenced by spatial position and salience of potential relata. Our results showed that spatial position indeed influenced reference production: relatum objects positioned on the left in the scene were more likely to be mentioned as (first) relatum than those positioned on the right. However, participants did not systematically opt for the leftmost relatum object, suggesting that there might be other factors that could influence reference production as well. Therefore, in Experiments 1b–d, we manipulated the (conceptual and perceptual) salience of relatum objects, and these manipulations had no effect. In particular, we did not find that relatum objects that were salient, because of animacy, by priming visual attention or by using salient colors, were more likely to be used as (first) relatum. In Experiment 2, we assess if spatial position and salience affect listeners' evaluations of spatial descriptions.

3. EXPERIMENT 2—LISTENER PREFERENCES

To further investigate the extent to which spatial position and salience might influence listeners' preferences for relata, in Experiment 2, participants were asked to rank relational descriptions. Given that many earlier studies have revealed strong effects of animacy, we expect descriptions that have an animate entity as (first) relatum to be ranked higher.

For pragmatic reasons, the language used in Experiment 2 was English. Earlier work on reference production (Theune et al., 2010; Koolen et al., 2012) suggested that English and Dutch are comparable in terms of the attributes used in descriptions.

3.1. Participants

Eighty-six English-speaking native participants from Australia, Canada and the UK were recruited via CrowdFlower, a crowdsourcing service similar to Amazon Mechanical Turk. The validity of this method for behavioral studies has been previously tested and studies assessing data quality have been positive about using crowdsourcing as an alternative to more traditional approaches of participant recruitment (e.g., Buhrmester et al., 2011; Crump et al., 2013). Ten participants' data were excluded for various reasons: because their ranking was identical (in more than 30% of the cases) to the order in which descriptions were presented (2 participants); because they declared being not native English speakers (5 participants); because did not finish the task (3 participants). The final sample included 66 participants (37 males, mean age 39.36 years, range 20–64 years).

3.2. Materials

The stimuli from Experiment 1b were used. The 32 experimental stimuli were divided across 6 randomized lists. The experiment consisted of 8 experimental stimuli (out of which 4 had an animate positioned left and 4 had an animate positioned right) and 8 filler stimuli. In addition, we used a set of four sentences representing the two participant formulation preferences using a single relatum and two relata. These sentences were translated from Dutch to English. The sentences were: *the ball in front of the ANIMATE* (e.g., the man); *the ball in front of the INANIMATE* (e.g., closet); *the ball between the ANIMATE and the INANIMATE*; *the ball between the INANIMATE and the ANIMATE.*

3.3. Procedure

First, participants were instructed to rank the four descriptions starting with the one they "liked best" given the visual scene.

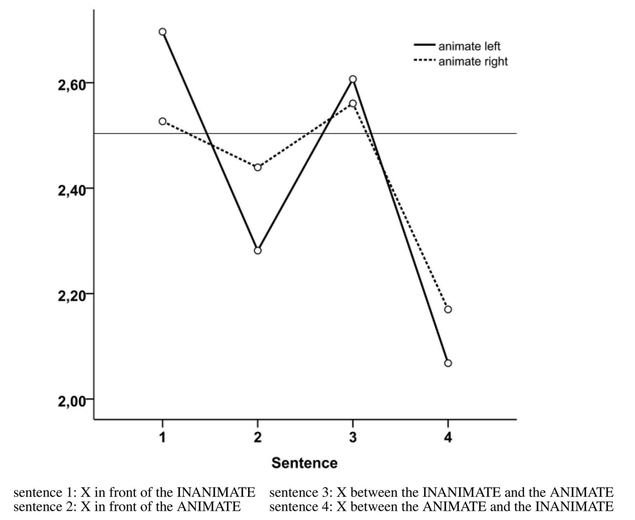


FIGURE 4 | Mean ranks across conditions (1 = highest preference, 4 = lowest preference), where 2.5 represents random chance.

The descriptions were presented under each scene in random order. The participant could rank the descriptions by dragging them in an input field with four empty slots, where the slot no. 1 represented the description that participants liked most, while slot no. 4 was assigned for the description that they liked least. The picture remained on the screen until the participants had made their choice and pressed a button to continue. Each experimental trial was followed by one filler trial.

3.4. Results and Discussion

For each trial, the order of the descriptions was ranked, starting from 1 (the best description) to 4 (the worst description).

Whether animacy influenced preferences was tested with a repeated measures ANOVA, having three within subjects factors: the Position of the Animate (2 levels: animate left, animate right), the Participant Formulation Preference (4 levels: in front of ANIMATE, in front of INANIMATE, between the ANIMATE and the INANIMATE, between the INANIMATE and the ANIMATE) and Scenes (4 levels)[2].

Results revealed a main effect of Participant Formulation Preference [$F_{(3,\ 306)} = 5.186$, $p = 0.002$, $\eta_p^2 = 0.048$] and a significant interaction between Animate Position and Participant Formulation Preference [$F_{(3,\ 306)} = 4.412$, $p = 0.005$, $\eta_p^2 = 0.041$]. Participants preferred the description that mentioned two relata and started with the animate irrespective of the visual scene (animate left $M = 2.07$, $SE = 0.11$; animate right $M = 2.17$ $SE = 0.11$) (see **Figure 4**). The second most preferred description was the one that mentioned a single relatum, namely the animate. This description was more preferred when the animate was positioned on the left of the scene ($M = 2.28$, $SE = 0.08$) than on the right of the scene [$M = 2.44$, $SE = 0.09$; $F_{(1,\ 102)} = 6.58$, $p = 0.003$, $\eta_p^2 = 0.082$]. The least preferred description was the one mentioning a single inanimate relatum, especially when the animate was placed on the left [$M = 2.70$, $SE = 0.09$; animate placed right $M = 2.53$, $SE = 0.09$; $F_{(1,\ 102)} = 9.08$, $p = 0.012$, $\eta_p^2 = 0.061$].

4. CONCLUSIONS AND DISCUSSION

The main aim of this study was to examine the extent to which production of spatial relational descriptions is influenced by spatial position and salience. Our results show that spatial position systematically influenced reference production. A basic language elicitation task determined that speakers often mentioned the entity positioned leftmost in the scene as (first) relatum. This was consistent across four production experiments (highest mean 67%, η_p^2 range 0.47–0.75). Based on these observations, we considered that other factors might influence reference production. Thus, we investigated possible effects of the objects' (conceptual and perceptual) salience. In Experiment 1b, conceptual salience was manipulated visually, by having an animate and an inanimate relatum candidate. Despite the strong body of research arguing for effects of animacy in reference production, animacy was found to have a significant effect in only one out of three production studies (Experiment 1d). Visual salience was manipulated using two different methods. In Experiment 1c, attention was primed using a flash and in Experiment 1d, the objects were made perceptually salient by having a distinctive color. These manipulations yielded no effects. From a listener's perspective, the formulation of the description and the position of the animate entity in the scene influenced to some extent the acceptability rating (Experiment 2). These results are further discussed in relation to broader aspects of reference production.

4.1. Relevance for Reference Production

The studies reported bring evidence for relatum reference being influenced by the inherent spatial structure of the scene, a factor largely unexplored in studies of (computational) reference production. Across different circumstances, there was a systematic preference for mentioning left entities as (first) relatum in relational descriptions such as *in front of X; in between X and Y*. This preference could have been caused either by cultural differences or spatial asymmetries in scene scanning. It is worth replicating Experiment 1 with speakers of a language with a right-to-left system.

The position of the object seems to be a constant factor influencing reference production. Our results are consistent with Miller et al. (2011), who stress that the spatial relation between the target and the relatum candidates is an important predictor in relatum selection. Congruent evidence comes from Clarke et al. (2013b), who report that position (measured in relation to the center of the screen) contributes to perceptual salience of the object and affects the likelihood with which objects are mentioned. When objects are symmetrically arranged, not only spatial position, but also salience influence (to some extent) referential choices.

Previous research has granted an important role to salience in reference production. Visually salient and linguistically important (e.g., animate) objects are more likely to be mentioned,

[2] The analyses were also done using non-parametric Friedman's signed rank tests which yielded similar results.

as well as objects spatially placed in a prominent position Clarke et al. (2013b). In these studies, we have manipulated salience on conceptual and visual levels. We expected salient entities to influence the ordering of linguistic elements in the spatial relation and be mentioned (first) more often than the other candidates. Surprisingly, there were poor effects of animacy, no effects of the visual salience manipulation. Below we address a few questions related to these results.

First, why did animacy have a limited influence on reference? The impact of animacy on word order, and more precisely on conjunctive phrases is debatable (see Branigan et al., 2008). For example, when the conjoined NPs are presented embedded in a sentence such as *the dog and the telephone were making noise* or *the surgeon yelled for a nurse and a needle* (experiments 1 and 2 in McDonald et al., 1993), animacy had no reliable effect on conjunct order. However, when removed from sentences and produced in isolated phrases (experiments 3, 4, and 5 in McDonald et al., 1993), animate nouns regularly occupied a leading position. It is conceivable that the effect of animacy in the current studies might have been dampened by sentence context, in line with the findings of McDonald et al. (1993). Compared to other experiments that found a strong effect of animacy on reference production in visual domains (e.g., Coco and Keller, 2009), in our studies animacy was manipulated visually, without priming participants with animacy in a lexical format. "Visual animacy" was suggested to be a less important factor in attention guiding (Wolfe and Horowitz, 2004). Interestingly, the results of the acceptability rating task (Experiment 2) present a different picture, which is more in line with previous studies suggesting strong effects of animacy and is in apparent contrast with the production data from Experiment 1. Descriptions which included an animate entity as the first (or the only) relatum were rated higher than those having an inanimate as first or single relatum. In fact, the descriptions which had animate as first relatum were rated as the most acceptable, irrespective of the spatial placement of the objects in the scene. Not only animacy, but also the left bias seemed to have influenced the acceptability ratings, as descriptions containing a single animate relatum, were rated higher when the animate entity was placed on the left, rather than on the right side of the visual scene and the same pattern was observed for descriptions that included a single inanimate relatum. This slight discrepancy between the results of Experiments 1, 2 highlights an observation that has been made before in the context of REG evaluation: what speakers do is not necessarily what is appreciated most by addressees (for a review, see Gatt and Belz, 2010; Krahmer and Van Deemter, 2012).

Second, why did priming attention have no effect? Directing speakers' attention to a specific region of the scene predicts which entity would be mentioned first, both in sentences and in conjoined NP descriptions (Gleitman et al., 2007). Yet, in our study, the attention capture cue did not influence utterances. Preference for left entities was stable, even when visual attention was directed to a different relatum candidate. It might be the case that the effect of the cue fades during production (the first-mentioned entity in our scenario was always the target ball). Other studies also report no effect of this attention priming manipulation (Nappa and Arnold, 2014; Arnold and Lao, 2015). In addition, when salience was explicitly manipulated by making an object perceptually salient, it did not yield a significant effect. This might be caused by the visual simplicity of the stimuli.

The extent to which our results can be observed using complex visual scenes also warrants further study. For example, Viethen and Dale (2008) reported (limited) effects of relatum salience in scenes consisting of three objects with simple spatial arrangements, but in a more complex study, salient large relata did not systematically influence whether the object was mentioned or not (Viethen et al., 2011). Similarly, participants describing routes through groups of colored objects in a MapTask (Louwerse et al., 2007) seem to have disregarded potential visual distractors (Viethen and Dale, 2011). The results of Elsner et al. (2014); Clarke et al. (2013a) reveal a different picture: in very cluttered and complex scenes, like the Where's Wally pictures, speakers were sensitive to perceptual salience, not only when choosing the objects to mention, but also when producing a description. The relational descriptions started more often with the salient object. Nonetheless, our studies are complementary, showing (though to a smaller extent) effects of the position an object occupies in the scene and salience.

Our experiments have a number of limitations. As mentioned above the scenes used as stimuli were simple and consisted of a small number of objects. Ideally, future research should take into account scenes of a higher visual complexity, use a different spatial arrangement of the objects and manipulate other perceptual features (such as size) as well. For a systematic analysis other tasks should be considered as well (e.g., testing listeners' comprehension in a reaction time study).

In the production experiment, we also discouraged participants from saying "the ball on the left." While objects in visual environments can be referred to with a wide variety of forms of spatial language, we wanted to focus on referential choices when describing objects in relations. However, we also acknowledge that identifying a target by mentioning its location (and thus, maybe contrasting the target with a potential distractor, see Tenbrink, 2005) is a widespread strategy. Crucially, more research is needed to find out when people need or prefer relational descriptions containing explicit relata.

4.2. Formulation Preferences

As for the formulations used, across studies, a small sample of participants chose a single relatum, thus producing a *X in front of Y* description. The chance of choosing one of the entities was not influenced by the distance between the relatum and the position of the distractor (the further away the relatum object was from the distractor ball, the less ambiguous).

Most of the participants referred to the target using the preposition *tussen* (in between), which describes the location of the target in relation to both relata. Compared with other locative prepositions, *in between* is a syntactically complex and cognitively more expensive one (because it contains more words and involves more relata), but it also provides a more accurate description. This preposition might be preferred due to the view point from which the speaker looks at the scene (Kelleher et al., 2010), from which the relatum candidates and the target seem arranged in an almost linear fashion. In fact, when the target object is situated between two other elements and the in between relation is available for reference, speakers will often use this option (Tenbrink, 2007, p.261).

4.3. Recommendations for Referring Expressions Algorithms

Understanding the criteria on which humans base their referential choices offers insights for the development of referring expressions generation algorithms. There are only few algorithms that make use of extrinsic attributes as a last resort (e.g., Dale and Haddock, 1991; Gardent, 2002; Krahmer and Theune, 2002; Krahmer et al., 2003; Varges, 2005). Crucially, more research is needed to find out when people need or prefer relational descriptions containing explicit relata. Nevertheless, these systems have little to say about relatum reference as they assume access to a predefined scene model, where the relata has been selected and treat spatial reference as the last means for generating a unique description. Though there are some assumptions regarding the factors that drive choices regarding relatum reference, there is no systematic research on this issue. For example, Krahmer and Theune (2002) note that human speakers and hearers might have a preference for relata which are close to the target. Kelleher et al. (2005) implement a measure for proximity and bring into discussion visual and discourse salience. Dos Santos Silva and Paraboni (2015) consider distance as the main factor, followed by the unique spatial relations between objects. Apart from distance, various other factors may influence relatum reference. For example, Elsner et al. (2014) highlight that visual features that contribute to the object's perceptual salience should be taken into account in order to generate more human-like reference in visual domains. Specifically, perceptual salience (spatial and visual information) influences the order in which relata are mentioned in relational descriptions.

Our results suggest that algorithms should take into account the spatial position and the object's salience. When the distance between target and the relatum candidates is similar, the spatial structure of the scene should be the first feature to be examined. In circumstances in which there are several relatum candidates similarly aligned, we suggest that entities placed on the left of the target to be favored. Perceptual and conceptual salience might also be taken into account. Given the practical nature of REG, the human-likeness aspect should be balanced with a comprehension-oriented perspective (e.g., Paraboni et al., 2007; Garoufi, 2013; Mast et al., 2014). Our results suggest that if the goal of the system is different from just producing a human-like expression, other factors might play a role (see also Krahmer and Van Deemter, 2012). More addressee oriented (and maybe more efficient) descriptions might be produced when including an animate as first relatum. Our results suggest that when the target object is situated between two other objects and the *in between* relation is available for reference, the system should refer to both objects and start with the animate irrespective of the position of the objects in the scene. However, if the system generates a description with a single relatum, this relatum preferably should be the object located on the left of the target.

Finally, speakers have to make several referential choices when uttering spatial descriptions and different factors can influence this process. The results of this study suggest that reference production was affected by the spatial position of a relatum candidate and less so by (conceptual and perceptual) salience.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

The first author received financial support from the Netherlands Organization for Scientific Research (NWO), Promoties in Geesteswetenschappen (322-89-008), which is greatly acknowledged.

ACKNOWLEDGMENTS

We are grateful to the Jugo research group for their useful comments on earlier versions of this paper. Preliminary data of this study were presented as poster at the PRE-CogSci 2013 Workshop on Production of referring expressions: Bridging the gap between cognitive and computational approaches to reference.

REFERENCES

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Arnold, J. E., and Lao, S.-Y. C. (2015). Effects of psychological attention on pronoun comprehension. *Lang. Cogn. Neurosci.* 30, 832–852. doi: 10.1080/23273798.2015.1017511

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011). Overspecification facilitates object identification. *J. Pragmat.* 43, 361–374. doi: 10.1016/j.pragma.2010.07.013

Barclay, M., and Galton, A. (2008). "An influence model for reference object selection in spatially locative phrases," in *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, eds C. Freksa, N. Newcombe, P. Gärdenfors, and S. Wölfl (Berlin: Springer), 216–232. doi: 10.1007/978-3-540-87601-4_17

Barclay, M., and Galton, A. (2013). "Selection of reference objects for locative expressions: the importance of knowledge and perception," in *Representing Space in Cognition: Interrelations o Behavior, Langauge, and Formal Models*, eds T. Tenbrink, J. Wiener, and C. Claramunt (Oxford, UK: Oxford University Press), 57–169.

Belke, E., and Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: analyses of viewing patterns and processing times during sameÌ-different decisions. *Eur. J. Cogn. Psychol.* 14, 237–266. doi: 10.1080/09541440143000050

Beun, R.-J., and Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmat. Cogn.* 6, 121–152. doi: 10.1075/pc.6.1-2.08beu

Bock, J. K., and Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21, 47–67. doi: 10.1016/0010-0277(85)90023-X

Bock, K., Loebell, H., and Morey, R. (1992). From conceptual roles to structural relations: bridging the syntactic cleft. *Psychol. Rev.* 99:150. doi: 10.1037/0033-295X.99.1.150

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *Pattern Anal. Mach. Intell. IEEE Trans.* 35, 185–207. doi: 10.1109/TPAMI.2012.89

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition* 75, 1–28. doi: 10.1016/S0010-0277(99)00073-6

Branigan, H. P., Pickering, M. J., and Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua* 118, 172–189. doi: 10.1016/j.lingua.2007.02.003

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). "Predicting the dative alternation," in *Cognitive Foundations of Interpretation*, eds G. Bouma, I. Kraemer, and J. Zwarts (Amsterdam: KNAW), 69–94.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

Butler, L. K., Tilbe, T. J., Jaeger, T. F., and Bohnemeyer, J. (2014). "Order of nominal conjuncts in visual scene description depends on language," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. eds P. Bello, M. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society), 284–289.

Carlson-Radvansky, L. A., Covey, E. S., and Lattanzi, K. M. (1999). What effects on where: functional influences on spatial relations. *Psychol. Sci.* 10, 516–521. doi: 10.1111/1467-9280.00198

Carlson-Radvansky, L. A., and Radvansky, G. A. (1996). The influence of functional relations on spatial term selection. *Psychol. Sci.* 7, 56–60. doi: 10.1111/j.1467-9280.1996.tb00667.x

Chan, T. T., and Bergen, B. (2005). "Writing direction influences spatial cognition," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, eds B. Bara, L. Barsalou, and M. Bucciarelli (Austin, TX: Cognitive Science Society), 412–417.

Chatterjee, A. (2001). Language and space: some interactions. *Trends Cogn. Sci* 5, 55–61. doi: 10.1016/S1364-6613(00)01598-9

Chokron, S., and De Agostini, M. (2000). Reading habits influence aesthetic preference. *Cogn. Brain Res.* 10, 45–49. doi: 10.1016/S0926-6410(00)00021-5

Chokron, S., and Imbert, M. (1993). Influence of reading habits on line bisection. *Cogn. Brain Res.* 1, 219–222. doi: 10.1016/0926-6410(93)90005-P

Clark, H. H., and Begun, J. S. (1971). The semantics of sentence subjects. *Lang. Speech* 14, 34–46.

Clarke, A. D., Elsner, M., and Rohde, H. (2013a). Where's wally: the influence of visual salience on referring expression generation. *Front. Psychol.* 4:329. doi: 10.3389/fpsyg.2013.00329

Clarke, A. D. F., Coco, M. I., and Keller, F. (2013b). The impact of attentional, linguistic and visual features during object naming. *Front. Psychol.* 4:927. doi: 10.3389/fpsyg.2013.00927

Coco, M. I., and Keller, F. (2009). "The impact of visual information on reference assignment in sentence production," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds N. Taatgen, H. van Rijn, L. Schomaker, and J. Nerbonne (Austin, TX: Cognitive Science Society), 274–279.

Coco, M. I., and Keller, F. (2015). Integrating mechanisms of visual guidance in naturalistic language production. *Cogn. Process.* 16, 131–150. doi: 10.1007/s10339-014-0642-0

Coventry, K. R., and Garrod, S. C. (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. Hove; New York, NY: Psychology Press; Taylor & Francis.

Craton, L. G., Elicker, J., Plumert, J. M., and Pick, H. L. (1990). Children's use of frames of reference in communication of spatial location. *Child Dev.* 61, 1528–1543. doi: 10.2307/1130762

Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE* 8:e57410. doi: 10.1371/journal.pone.0057410

Dale, R., and Haddock, N. (1991). "Generating referring expressions involving relations," in *Proceedings of the Fifth Conference of the European Association for Computational Linguistics*, eds J. Kunze, and D. Reimann (Berlin: Association for Computational Linguistics), 161–166. doi: 10.3115/977180.977208

Dale, R., and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902_3

Davies, C., and Katsos, N. (2009). "Are interlocutors as sensitive to over-informativeness as they are to under-informativeness," in *Proceedings of the Pre-CogSci Workshop on the Production of Referring Expressions* (Amsterdam).

de Vega, M., Rodrigo, M. J., Ato, M., Dehn, D. M., and Barquero, B. (2002). How nouns and prepositions fit together: an exploration of the semantics of locative sentences. *Discourse Process.* 34, 117–143. doi: 10.1207/S15326950DP3402_1

Dickinson, C. A., and Intraub, H. (2009). Spatial asymmetries in viewing and remembering scenes: consequences of an attentional bias? *Atten. Percep. Psychophys.* 71, 1251–1262. doi: 10.3758/APP.71.6.1251

Dos Santos Silva, D., and Paraboni, I. (2015). Generating spatial referring expressions in interactive 3d worlds. *Spatial Cogn. Comput.* 15, 186–225. doi: 10.1080/13875868.2015.1039166

Downing, P. E., Bray, D., Rogers, J., and Childs, C. (2004). Bodies capture attention when nothing is expected. *Cognition* 93, 27–38. doi: 10.1016/j.cognition.2003.10.010

D'Zmura, M. (1991). Color in visual search. *Vision Res.* 31, 951–966.

Elsner, M., Rohde, H., and Clarke, A. D. (2014). "Information structure prediction for visual-world referring expressions," in *14th Conference of the European Chapter of the Association for Computational Linguistics*, eds S. Wintner, S. Goldwater, and S. Riezler (Gothenburg), 520–530. doi: 10.3115/v1/E14-1055

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., and Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception* 37, 571–583. doi: 10.1068/p5705

Folk, C. L., Remington, R. W., and Wright, J. H. (1994). The structure of attentional control: contingent attentional capture by apparent motion, abrupt onset, and color. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 317–329. doi: 10.1037/0096-1523.20.2.317

Forrest, L. B. (1996). "Discourse goals and attentional processes in sentence production: the dynamic construal of events," in *Conceptual Structure, Discourse and Language*, ed A. Goldberg (Stanford, CA. Center for the Study of Language and Information), 149–161.

Foulsham, T., Gray, A., Nasiopoulos, E., and Kingstone, A. (2013). Leftward biases in picture scanning and line bisection: a gaze-contingent window study. *Vision Res.* 78, 14–25. doi: 10.1016/j.visres.2012.12.001

Fukumura, K., and van Gompel, R. P. (2011). The effect of animacy on the choice of referring expression. *Lang. Cogn. Process.* 26, 1472–1504. doi: 10.1080/01690965.2010.506444

Gardent, C. (2002). "Generating minimal definite descriptions," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ed J. Stephen (Philadelphia: Association for Computational Linguistics), 96–103.

Garoufi, K. (2013). *Interactive Generation of Effective Discourse in Situated Context: A Planning-based Approach*. Ph.D thesis, University of Postdam.

Gatt, A., and Belz, A. (2010). "Introducing shared tasks to NLG: The TUNA shared task evaluation challenges," in *Empirical Methods in Natural Language Generation*, eds E. Krahmer, and M. Theune (Berlin; Heidelberg: Springer-Verlag), 264–293.

Gatt, A., Krahmer, E., van Deemter, K., and van Gompel, R. P. (2014). Models and empirical data for the production of referring expressions. *Lang. Cogn. Neurosci.* 29, 899–911. doi: 10.1080/23273798.2014.933242

Gentner, D., Özyürek, A., Gürcanli, Ö., and Goldin-Meadow, S. (2013). Spatial language facilitates spatial cognition: evidence from children who lack language input. *Cognition* 127, 318–330. doi: 10.1016/j.cognition.2013.01.003

Gleitman, L. R., January, D., Nappa, R., and Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *J. Mem. Lang.* 57, 544–569. doi: 10.1016/j.jml.2007.01.007

Goudbeek, M., and Krahmer, E. (2012). Alignment in interactive reference production: content planning, modifier ordering, and referential overspecification. *Topics Cogn. Sci.* 4, 269–289. doi: 10.1111/j.1756-8765.2012.01186.x

Henderson, J., and Ferreira, F. (2013). *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. New York, NY: Psychology Press.

Hund, A. M., and Plumert, J. M. (2007). What counts as by? young children's use of relative distance to judge nearbyness. *Dev. Psychol.* 43, 121. doi: 10.1037/0012-1649.43.1.121

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203. doi: 10.1038/35058500

Jaeger, T. F., and Tily, H. (2011). On language utility: processing complexity and communicative efficiency. *Wiley Interdiscipl. Rev. Cogn. Sci.* 2, 323–335. doi: 10.1002/wcs.126

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (2014). "Referitgame: referring to objects in photographs of natural scenes," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, ed A. Moschitti (Stroudsburg: Association for Computational Linguistics), 787–798. doi: 10.3115/v1/D14-1086

Kelleher, J., Costello, F., and van Genabith, J. (2005). Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artif. Intell.* 167, 62–102. doi: 10.1016/j.artint.2005.04.008

Kelleher, J. D., Ross, R. J., Mac Namee, B., and Sloan, C. (2010). *Situating Spatial Templates for Human-Robot Interaction.* Arlington, Virginia: American Association for Artificial Intelligence Fall Symposium Series.

Kirchner, H., and Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Res.* 46, 1762–1776. doi: 10.1016/j.visres.2005.10.002

Kirsh, D. (1995). The intelligent use of space. *Artif. Intell.* 73, 31–68. doi: 10.1016/0004-3702(94)00017-U

Kollmorgen, S., Nortmann, N., Schröder, S., and König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Comput. Biol.* 6:e1000791. doi: 10.1371/journal.pcbi.1000791

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Koolen, R., Krahmer, E., and Theune, M. (2012). "Learning preferences for referring expression generation: effects of domain, language and algorithm," in *Proceedings of the Seventh International Natural Language Generation Conference*, eds B. Di Eugenio, and S. McRoy (Stroudsburg: Association for Computational Linguistics), 3–11.

Krahmer, E., and Theune, M. (2002). "Efficient context-sensitive generation of referring expressions," in *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, eds K. van Deemter, and R. Kibble (Stanford: CSLI Publications), 223–263.

Krahmer, E., and Van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI_a_00088

Krahmer, E., Van Erk, S., and Verleg, A. (2003). Graph-based generation of referring expressions. *Comput. Linguist.* 29, 53–72. doi: 10.1162/089120103321337430

Lakoff, G., and Johnson, M. (2008). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Levelt, W. J. (1993). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levinson, S. C. (1996). "Frames of reference and Molyneux's question: cross-linguistic evidence," in *Language and Space*, eds P. Bloom, M. Peterson, L. Nadel, and M. Garrett (Cambridge, MA: MIT Press), 109–169.

Levinson, S. C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge, MA: Cambridge University Press.

Louwerse, M., Benesh, N., Hoque, M., Jeuniaux, P., Lewis, G., Wu, J., and Zirnstein, M. (2007). "Multimodal communication in face-to-face conversations," in *Proceedings of the 29th Annual Cognitive Science Society*, eds D. S. McNamara, and J. G. Trafton (Austin, TX: Cognitive Science Society), 1235–1240.

Maass, A., and Russo, A. (2003). Directional bias in the mental representation of spatial events nature or culture? *Psychol. Sci.* 14, 296–301. doi: 10.1111/1467-9280.14421

Mast, V., Couto Vale, D., and Falomir, Z. (2014). "Enabling grounding dialogues through probabilistic reference handling," in *Proceedings of RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, eds A. Eshghi, K. Fukumura, and S. Janarthanam (Edinburgh).

McDonald, J. L., Bock, K., and Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cogn. Psychol.* 25, 188–230.

Meyer, A. S., Sleiderink, A. M., and Levelt, W. J. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition* 66, 25–33. doi: 10.1016/S0010-0277(98)00009-2

Miller, J. E., Carlson, L. A., and Hill, P. L. (2011). Selecting a reference object. *J. Exp. Psychol. Learn. Mem. Cogn.* 37, 840–850. doi: 10.1037/a0022791

Moratz, R., and Tenbrink, T. (2006). Spatial reference in linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations. *Spatial Cogn. Comput.* 6, 63–107. doi: 10.1207/s15427633scc0601_3

Myachykov, A., Thompson, D., Scheepers, C., and Garrod, S. (2011). Visual attention and structural choice in sentence production across languages. *Lang. Linguist. Compass* 5, 95–107. doi: 10.1111/j.1749-818X.2010.00265.x

Nappa, R., and Arnold, J. E. (2014). The road to understanding is paved with the speakers intentions: cues to the speakers attention and intentions affect pronoun comprehension. *Cogn. Psychol.* 70, 58–81. doi: 10.1016/j.cogpsych.2013.12.003

New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proc. Natl. Acad. Sci. U.S.A.* 104, 16598–16603. doi: 10.1073/pnas.0703913104

Onishi, K. H., Murphy, G. L., and Bock, K. (2008). Prototypicality in sentence production. *Cogn. Psychol.* 56, 103–141. doi: 10.1016/j.cogpsych.2007.04.001

Ossandón, J. P., Onat, S., and König, P. (2014). Spatial biases in viewing behavior. *J. Vis.* 14:20. doi: 10.1167/14.2.20

Paraboni, I., and van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Lang. Cogn. Neurosci.* 29, 1002–1017. doi: 10.1080/01690965.2013.805796

Paraboni, I., Van Deemter, K., and Masthoff, J. (2007). Generating referring expressions: making referents easy to identify. *Comput. Linguist.* 33, 229–254. doi: 10.1162/coli.2007.33.2.229

Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Res.* 42, 107–123. doi: 10.1016/S0042-6989(01)00250-4

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 89–110. doi: 10.1515/ling.1989.27.1.89

Prat-Sala, M., and Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: a cross-linguistic study in English and Spanish. *J. Mem. Lang.* 42, 168–182. doi: 10.1006/jmla.1999.2668

Talmy, L. (2003). *Toward a Cognitive Semantics*, *Vol. 1*. Cambridge, MA: MIT Press.

Tanaka, J., Weiskopf, D., and Williams, P. (2001). The role of color in high-level vision. *Trends Cogn. Sci.* 5, 211–215. doi: 10.1016/S1364-6613(00)01626-0

Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: reinterpreting salience. *J. Vis.* 11:5. doi: 10.1167/11.5.5

Tatler, B. W., and Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Vis. Cogn.* 17, 1029–1054. doi: 10.1080/13506280902764539

Taylor, H. A., and Rapp, D. N. (2004). Where is the donut? factors influencing spatial reference frame use. *Cogn. Process.* 5, 175–188. doi: 10.1007/s10339-004-0022-2

Taylor, H. A., and Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *J. Mem. Lang.* 31, 261–292. doi: 10.1016/0749-596X(92)90014-O

Taylor, T. E., Gagné, C. L., and Eagleson, R. (2000). "Cognitive constraints in spatial reasoning: reference frame and reference object selection," in *American Association for Artificial Intelligence Technical Report SS-00-04*, eds A. Butz, A. Krger, and P. Olivier (Menlo Park, CA: Association for the Advancement of Artificial Intelligence Press), 168–172.

Tenbrink, T. (2005). "Identifying objects on the basis of spatial contrast: an empirical study," in *Spatial Cognition IV. Reasoning, Action, Interaction*, eds C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, and T. Barkowsky (Berlin: Springer), 124–146. doi: 10.1007/978-3-540-32255-9_8

Tenbrink, T. (2007). *Space, Time, and the Use of Language: An Investigation of Relationships*. Berlin: Walter de Gruyter.

Tenbrink, T. (2011). Reference frames of space and time in language. *J. Pragmat.* 43, 704–722. doi: 10.1016/j.pragma.2010.06.020

Theeuwes, J. (1994). Endogenous and exogenous control of visual selection. *Perception* 23, 429–440. doi: 10.1068/p230429

Theune, M., Koolen, R., and Krahmer, E. (2010). "Cross-linguistic attribute selection for reg: Comparing Dutch and English," in *Proceedings of the 6th International Natural Language Generation Conference*, eds J. Kelleher, B. Mac Namee, and I. van der Sluis (Stroudsburg: Association for Computational Linguistics), 191–195.

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Tversky, B. (2011). Visualizing thought. *Topics Cogn. Sci.* 3, 499–535. doi: 10.1111/j.1756-8765.2010.01113.x

Tversky, B., Kugelmass, S., and Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cogn. Psychol.* 23, 515–557. doi: 10.1016/0010-0285(91)90005-9

Tversky, B., Lee, P., and Mainwaring, S. (1999). Why do speakers mix perspectives? *Spatial Cogn. Comput.* 1, 399–412. doi: 10.1023/A:1010091730257

van Deemter, K., Gatt, A., van Gompel, R. P., and Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01187.x

Varges, S. (2005). "Spatial descriptions as referring expressions in the maptask domain," in *Proceedings of the 10th European Workshop on Natural Language Generation*, eds A. K. J. Graham Wilcock, C. Mellish, and E. Reiter (Scotland), 207–210.

Viethen, J., and Dale, R. (2008). "The use of spatial relations in referring expression generation," in *Proceedings of the Fifth International Natural Language Generation Conference*, eds M. White, C. Nakatsu, and D. McDonald (Stroudsburg: Association for Computational Linguistics), 59–67. doi: 10.3115/1708322.1708334

Viethen, J., and Dale, R. (2010). "Speaker-dependent variation in content selection for referring expression generation," in *Proceedings of the 8th Australasian Language Technology Workshop*, eds I. Nitin, and Z. Simon (Australasian Language Technology Association), 81–89.

Viethen, J., and Dale, R. (2011). "Gre3d7: a corpus of distinguishing descriptions for objects in visual scenes," in *Proceedings of the UCNLG and Eval: Language Generation and Evaluation Workshop*, eds A. Belz, R. Evans, A. Gatt, and K. Striegnitz (Edinburgh: Association for Computational Linguistics), 12–22.

Viethen, J., Dale, R., and Guhe, M. (2011). "The impact of visual context on the content of referring expressions," in *Proceedings of the 13th European Workshop on Natural Language Generation*, eds G. Claire and S. Kristina (Stroudsburg: Association for Computational Linguistics), 44–52.

Vogels, J., Krahmer, E., and Maes, A. (2013). When a stone tries to climb up a slope: the interplay between lexical and perceptual animacy in referential choices. *Front. Psychol.* 4:154. doi: 10.3389/fpsyg.2013.00154

Vorwerg, C., and Tenbrink, T. (2007). "Discourse factors influencing spatial descriptions in English and German," in *Spatial Cognition V. Reasoning, Action, Interaction*, eds T. Barkowsky, M. Knauff, G. Ligozat, and D. R. Montello (Berlin; Heidelberg: Springer-Verlag), 470–488. doi: 10.1007/978-3-540-75666-8_27

Westerbeek, H., Koolen, R., and Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Front. Psychol.* 6:935. doi: 10.3389/fpsyg.2015.00935

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision Res.* 34, 1187–1195. doi: 10.1016/0042-6989(94)90300-X

Wolfe, J. M., and Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nat. Rev. Neurosci.* 5, 495–501. doi: 10.1038/nrn1411

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 09 December 2015
doi: 10.3389/fpsyg.2015.01793

Giving Good Directions: Order of Mention Reflects Visual Salience

Alasdair D. F. Clarke[1]*, *Micha Elsner*[2] *and Hannah Rohde*[3]

[1] School of Psychology, The College of Life Sciences and Medicine, University of Aberdeen, Aberdeen, UK, [2] Department of Linguistics, The Ohio State University, Columbus, OH, USA, [3] Linguistics and English Language, University of Edinburgh, Edinburgh, UK

In complex stimuli, there are many different possible ways to refer to a specified target. Previous studies have shown that when people are faced with such a task, the content of their referring expression reflects visual properties such as size, salience, and clutter. Here, we extend these findings and present evidence that (i) the influence of visual perception on sentence construction goes beyond content selection and in part determines the order in which different objects are mentioned and (ii) order of mention influences comprehension. Study 1 (a corpus study of reference productions) shows that when a speaker uses a relational description to mention a salient object, that object is treated as being in the common ground and is more likely to be mentioned first. Study 2 (a visual search study) asks participants to listen to referring expressions and find the specified target; in keeping with the above result, we find that search for easy-to-find targets is faster when the target is mentioned first, while search for harder-to-find targets is facilitated by mentioning the target later, after a landmark in a relational description. Our findings show that seemingly low-level and disparate mental "modules" like perception and sentence planning interact at a high level and in task-dependent ways.

Keywords: referring expressions, visual search, visual salience

OPEN ACCESS

Edited by:
Albert Gatt,
University of Malta, Malta

Reviewed by:
Alfons Maes,
Tilburg University, Netherlands
Ivandre Paraboni,
University of Sao Paulo, Brazil

****Correspondence:***
Alasdair D. F. Clarke
a.clarke@abdn.ac.uk

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *12 August 2015*
Accepted: *06 November 2015*
Published: *09 December 2015*

Citation:
Clarke ADF, Elsner M and Rohde H (2015) Giving Good Directions: Order of Mention Reflects Visual Salience. Front. Psychol. 6:1793. doi: 10.3389/fpsyg.2015.01793

1. INTRODUCTION

When referring to an entity (the *target*) in a visual scene, speakers often describe it relative to some nearby *landmark*: "the woman next to the stairs." Previous research demonstrates that speakers choose these landmarks with reference to the visual properties of the scene, and in particular that they prefer those that are larger and easier to see (Kelleher et al., 2005; Duckham et al., 2010; Clarke et al., 2013). Much less is known about how these perceptual effects extend to the information-structural ordering of elements in a description. Although alternative orders are available ("next to the stairs is a woman"), most existing models of reference do not address the production format question: how speakers choose to package the content of a referring expression when it includes both a target and one or more disambiguating landmarks. In this work, we demonstrate via a corpus study of reference productions that visual perception influences the order chosen: larger and more visually salient landmarks are more likely to precede the target. The results from a subsequent comprehension study using a visual search task show that this pattern of ordering also helps the listener to find the target faster. The production and comprehension results indicate that dialogue participants' perceptions of the scene have far-reaching effects on both referring expression generation (REG) and understanding. Visual perception is not confined to providing inputs to a content selection mechanism, as in many popular models, but also contributes toward high-level decisions about the expression's structure.

Theories which acknowledge a role for perception in ordering the description do so in two ways. In least-effort theories, speakers compose references using cognitively inexpensive heuristics (Beun and Cremers, 1998). In particular, speakers order large objects first because they see them earliest. Such an approach is in line with egocentric models of production in which speakers use what they are familiar with to estimate what objects may be visible and shared (Horton and Keysar, 1996). Neo-Gricean theories, on the other hand, treat ordering preferences as an example of *audience design*, in which speakers construct referring expressions which will help their listeners find the target quickly and easily. Thus, one critical prediction of the neo-Gricean approach is that such speaker behavior is actually helpful for listeners.

Our visual search study shows that this is in fact the case: listeners find the target object faster when a highly salient landmark is referred to earlier rather than later, and when a difficult-to-see landmark is referred to later rather than earlier. Thus, neo-Gricean theories remain a viable explanation for the ordering preference. In particular, the pattern fits neatly into more general theories of *information structure* which state that given (familiar) information typically precedes new information in the sentence (Prince, 1981; Ward and Birner, 2001). Although many researchers have stated that perceptually salient entities can be treated as familiar by discourse participants (Ariel, 1988; Roberts, 2003), few have given a detailed account of the kinds of perceptual factors which contribute. Cognitive semantics defines partitions in cognitive semantics between figure and ground (Talmy, 1978): Figures are elements that are smaller or less immediately perceivable (visual salience) and of greater concern or relevance (task salience), while Ground is likely to be larger, more immediately perceivable, and more familiar. Although the work on figure and ground indicates how elements in complex descriptions relate, it does not specify which orderings are preferred in production or comprehension. Here we show, in line with prior work on information structure and the on distinction between figure and ground, that computational models of visual salience correctly predict which objects speakers are likely to place earlier in their descriptions. Furthermore, listeners are found to be sensitive to order of mention, showing facilitation when a target that is easy to find is mentioned first and also when a hard-to-find target is preceded by a mention of a more salient easy-to-find landmark.

Earlier studies evaluating automatically generated referring expressions have shown that the most human-like ones are not always the most helpful for listeners (Belz and Gatt, 2008), suggesting that at least some tendencies in human REG do not involve clear estimates of listener needs. Our results imply that information structural patterns are not among them, and on the contrary may even be the product of deliberate optimization. Moreover, although systems for automatic REG have given little attention to ordering in the past, our results suggest that the use of perceptual data may lead to both more human-like references and better performance.

2. MOTIVATION

Humans are highly proficient at REG, and human-like performance is often taken as a goal for automatic REG systems (Viethen and Dale, 2006). But more human-like referring expressions are not necessarily more helpful ones. Large individual differences are often found in RE production, and it is reasonable to expect that some speakers will be better at giving good instructions than others. Belz and Gatt (2008) compare task-based evaluations (search time and accuracy) to intrinsic ones (string similarity to human models) on computationally generated referring expressions from the ASGRE challenge (Belz and Gatt, 2007) and find no correlation between the two. While this experiment involved simple domains (furniture and people, identified by discrete-valued attributes), it stands as a warning that not all human behavior in REG should be interpreted as facilitating visual search. Thus, the question of ordering preferences for relative descriptions is really two questions: how speakers actually behave, and how they should normatively behave to facilitate visual search for listeners.

REG models which use relative descriptions are often separated into those focused on *identifying* a target object among distractors and those *locating* it in space (Barclay, 2010). We view both of these as strategies for accomplishing the higher-level goal of placing an unknown but visible entity into *common ground* (Clark and Wilkes-Gibbs, 1986), the set of entities which each participant knows is familiar to the other. However, the properties of the domain and task constraints may affect which of these strategies is most appropriate, and therefore what sort of behavior experimenters observe.

In relatively small domains where targets are easy to spot, the primary focus is on identification. When human speakers generate relative descriptions for easy-to-see targets, they mention the landmark after the target, as in the GRE3D7 corpus (Viethen and Dale, 2011), which was specifically set up to elicit relative descriptions using small 3-dimensional images of geometric objects. Models of REG in these kinds of domains (surveyed in Krahmer and van Deemter, 2012) do not emphasize ordering strategies or the need to make syntactic decisions during the planning phase.

Models for visually complex domains such as direction-giving (Barclay, 2010; Gkatzia et al., 2015) must both disambiguate and locate the target. Even when the target is unambiguous, it may still be necessary to use disambiguating descriptions for landmarks (Barclay, 2010). Studies in this kind of domain have followed Talmy (1983) in finding that large, relatively stationary "background" objects make good landmarks for locating an entity rather than simply disambiguating it. For the most part, however, these studies have also focused on what is said (the choice of landmarks and prepositions) rather than the order of mention and the syntactic strategies used to achieve it.

This study extends an earlier one, Elsner et al. (2014), which does look for ordering preferences in human-authored relative descriptions. That study found that larger objects were more likely to be ordered earlier in the description. However, there was no effect on order of mention from a low-level visual salience model, raising potential doubts about whether ordering preferences are truly driven by visual salience. The lack of effect for salience could potentially be due to poor performane of the computational visual salience models: many different salience models have been developed over the last 15 years and there is no agreed on standard, or even a strict defintion of what is meant by low-level salience! Furthermore, our stimuli consisted of cluttered cartoon images which may be problomatic for models trained on photographs of natural scenes. In this study, we re-analyze the same data with a more sophisticated salience model and obtain an improved fit to the data, suggesting that the hypothesized effect of low-level salience is real. Duan et al. (2013), studying the same corpus, find visual effects on determiner selection, and similarly conclude that perception has an impact on late stages of the generation pipeline. These studies focus on generation, leaving open the question of whether the effects they observed were useful to listeners or not.

The question of which speaker behaviors help listeners is tightly connected to the question of whether speakers actively reason about their audience to *try* to help them, a process called *audience design*. Experimental evidence for audience design is widespread. Speakers overspecify descriptions more when they believe the task is important (for example, instructing a surgeon on which tool to use; Arts et al., 2011). They can keep track of which objects they've discussed with a particular listener (Horton and Gerrig, 2002). And they are more likely to tell listeners about an atypical element of an illustrated action ("stabbed with an icepick" vs. "a knife") when they know listeners can't see the illustration (Lockridge and Brennan, 2002). Audience design is widely accepted as a theoretical assumption underlying neo-Gricean models of reference (Frank and Goodman, 2012; Vogel et al., 2013) and experiments with language games (Degen and Franke, 2012; Rohde et al., 2012). But despite speakers' capabilities for design, not all speaker behavior is audience-driven. Speakers also try to minimize their own effort by mentioning objects and attributes in the order they see them (Pechmann, 1989), avoiding cognitively expensive scanning of irrelevant parts of the scene (Beun and Cremers, 1998), and using their own private knowledge as a proxy for common ground (Horton and Keysar, 1996). Strategies like these make the speaker's task easier, but these savings potentially come at the listener's expense.

Both models offer potential explanations for order-of-mention effects. Pechmann (1989) describes speakers' use of non-canonical adjective orders ("red big") for visual scenes and argues that such orderings result from an incremental sentence planning strategy (speakers initially perceive the target object's color and only later establish its size relative to other objects in the scene).

Accounts of ordering preferences in non-visual settings usually attribute them to audience design in the form of information-structural principles. Prince (1981) distinguishes between entities which are new to the discourse and those which have previously been mentioned. The first element in an English sentence is generally reserved for old information (already in common ground), while new information is placed at the end (Ward and Birner, 2001, *inter alia*). A variety of non-canonical syntactic constructions, such as *there*-insertion, are analyzed as strategies for enforcing these structural principles. In particular, Maienborn (2001) states that sentence-initial locatives can be *frame-setting* modifiers, which are a type of sentence topic explaining in what context the remaining information is to be interpreted. Information-structural ordering principles can be said to be driven by audience design, since understanding what information is in common ground requires reasoning about the listener. In particular, objects which are clearly perceptually accessible to the listener are treated as familiar (Roberts, 2003).

Thus, the ordering preferences examined here could arise from either mechanism. In an effort-minimization model, speakers talk earlier about large objects because they notice them first. In an audience-design model, speakers talk earlier about large objects because they believe their listeners will notice them first. Thus, either model predicts that more visually salient objects are placed early in the sentence. Our first contribution is to verify that this prediction is in fact true.

The two models differ in their predictions about listener behavior. If the ordering effect is due to effort minimization, it may or may not be helpful for listeners. If it is due to audience design, then (assuming speakers who try to be helpful actually are so), it should facilitate listeners' visual search for the target. Thus, if this ordering principle does not facilitate visual search, it cannot be an audience design effect. Our second contribution is to show that it does in fact facilitate visual search.

3. CORPUS STUDY

In this section, we test whether speakers prefer to place visually salient landmarks earlier in their referring expressions. The study expands upon Elsner et al. (2014), which used the same corpus of referring expressions, by adding better models of low-level visual salience in order to demonstrate that the effect is actually salience-driven, and includes an additional feature that encodes whether the landmark is spatially located to the left or right of the target in the scene. The procedures for using mixed-effects linear models have also been altered slightly in line with recommendations by Barr et al. (2013).

A relative description of an object has two elements: the *anchor* (the object to be located) and the *landmark* (mentioned only as an aid). Typically the anchor is the *target* of the expression overall, but some REs nest relative descriptions— "the woman next to the man next to the building"— in which case "man" is the landmark relative to "woman" but the anchor relative to "building."

In a complex image like the scenes in Where's Wally (see **Figure 1**), there are many ways to describe a particular entity. We distinguish four strategies for ordering the landmark relative to the anchor, which we illustrate with examples from our corpus (all referring to targets in **Figure 1**), with text describing the landmark in *italics* and text describing the anchor (in these cases also the target) in bold:

- PRECEDE: Directly in front of *the crypt that is green* there is **a man with no shirt and a white wrap on**.
- PRECEDE-ESTABLISH: Find *the sphinx (half man half lion)*. To the left of *it* is **a guy holding a red vase with a stripe on it**.

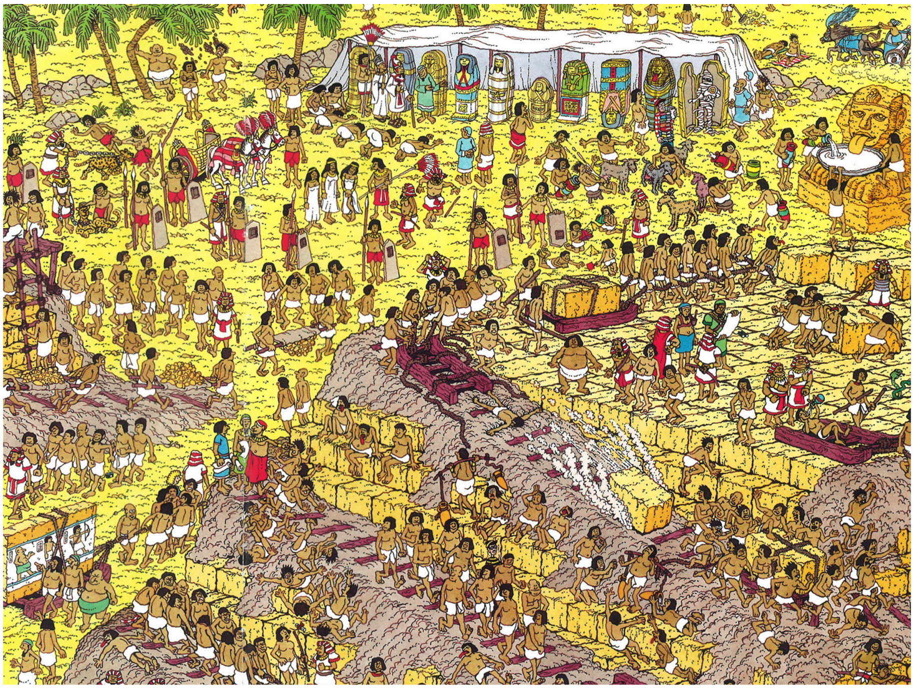


FIGURE 1 | Example stimulus used in the production and comprehension studies. In production, participants had to identify a designated target. In comprehension, the four referring expressions for this trial were (i) "at the upper right, the sphinx" [landmark only]; (ii) "at the upper right, the man holding the red vase with a stripe" [target only]; (iii) "at the upper right, the man holding the red vase with a stripe to the left of the sphinx" [landmark follows target]; (iv) "at the upper right, to the left of the sphinx, the man holding the red vase with a stripe on it" [landmark precedes target].

- INTERLEAVED: Near the bottom right, **a man walking** beside *the rock* **with his right foot forward**.
- FOLLOW: **The man in a white loincloth** at the upper left of the picture **standing** next to *a bald man*.

These ordering strategies[1] are distinguished based on the surface order of first mentions in the text. In the PRECEDE strategy, the first mention of the landmark occurs before any mention of the anchor. In the PRECEDE-ESTABLISH strategy, the landmark is first mentioned in its own clause, without a relation to the anchor (typically using "there is," "look," or "find"), and related to the anchor later. In the INTERLEAVED strategy, the anchor is described first, then the landmark, and then the anchor again. In the FOLLOW strategy, the anchor is mentioned first, then the landmark.

3.1. Dataset and Annotation

We analyze a collection of referring expressions for target people in images taken from the Where's Wally childrens picture books (Handford, 1987, 1988, 1993). The dataset[2] was originally collected by Clarke et al. (2013) in a study showing the effects of perceptual features (clutter and salience) on the selection of landmarks in REs. Mechanical Turk was used to collect the data using a task in which participants were asked to produce descriptions for targets over 11 images. In each image, 16 cartoon people were designated as targets and each participant saw each scene only once, with one of the targets designated with a colored box, as shown in **Figure 1**. The participant was instructed to type a description of the person in the box so that another person viewing the same scene (but without the box) would be able to find them.

The text of the instructions is shown in **Figure 2**. It asks participants to both identify and locate the target object (and as such is conceptually similar to the "please, pick up the X" frame used in Viethen and Dale, 2011).

Participants were trained on what makes a good referring expression in this domain by carrying out two visual searches based on different descriptions. The dataset contains 1672 descriptions, contributed by 152 different participants.

The REs are annotated for visual and linguistic content. The annotation scheme indicates which substrings of the RE describe the target object, another mentioned object or an image region such as "the left of the picture." References to parts or attributes of objects are not treated as separate objects; "a man holding a red vase" in **Figure 1** is a single object. The mentioned objects are linked to bounding boxes (or for very large objects, bounding polygons) in the image. For each mention of a non-target object, the annotation indicates whether it is part of a relational description of a specific anchor, and if so which; if it is not, it receives an ESTABLISH tag. These annotations are used

[1] There are also six examples of ESTABLISH constructions without the PRECEDE order, which we discard from further analysis.

[2] Released as the Wally Referring Expressions Corpus (WREC): http://datashare.is.ed.ac.uk/handle/10283/337.

You will see a series of pictures (30 in total). In each picture, there will be one person who is marked with a superimposed circle. Your task is to write a description of that person, such that someone else reading your description and seeing the same picture without the superimposed circle would be able to identify which person you intended.

- Your description should make it possible to identify the intended person quickly and easily.
- Give as much or as little detail as you think will help.
- Treat each picture as a separate item.

FIGURE 2 | Instructions for the picture description task in Clarke et al. (2013).

to determine the ordering strategies used in this study. In some cases, the linkage between objects is implicit:

- … *a group of 11 slaves is following* a slavemaster from left to right across the image. Choose **the third slave in line (the second bald slave)** [=of the 11 slaves].

In the RE above, the "group of 11 slaves" is introduced with an ESTABLISH construction, since in that clause, the group is not used as a landmark to locate another object. The group is later used as a landmark (implicitly, via the expression "third slave"). Since the first mention of the group precedes the anchor "third slave," this is marked PRECEDE, and therefore falls into the PRECEDE-ESTABLISH pattern.

TABLE 1 | One-vs.-all regression effects predicting order of anchor and landmark in relative descriptions.

	PRECEDE	PRECEDE-EST	INTER	FOLLOW
% (*n*) Instances	28% (918)	15% (493)	24% (797)	33% (1081)
intercept	2.64	−3.38	−2.44	−5.26
anch area	−0.42**	−0.21	−0.22**	0.40**
anch centr	0.16*	X	X	−0.13
anch deps	−0.19	−0.77**	0.26**	0.11
anch=targ	0.16	−0.32	0.84**	−0.80**
anch sal	−0.09	0.00	0.00	0.05
distance	0.02	X	X	0.03
sign. lr. dist.	−0.01	X	X	0.01
lmk=reg	15.68**	$-\infty$	$-\infty$	−16.42**
lmk area	3.97**	−0.67	1.53**	−4.48**
lmk centr	−1.12**	−1.03	−0.03	1.37**
lmk deps	0.07	1.31**	−0.57**	−0.75**
lmk sal	0.22**	0.13	−0.07	−0.17*

3.2. Distribution of Ordering Strategies

Our analysis covers each pair of anchor and landmark mentioned in the corpus (often more than one per description). In all, there are 3290 such pairs in the dataset. As shown in the first row of **Table 1**, the PRECEDE strategies, in aggregate, slightly outnumber the FOLLOW strategy; this is due to the overwhelming preference for image regions ("the left") to precede their anchors. The INTERLEAVED ordering is less common, but still quite well-represented.

To verify that this distribution does not simply reflect different participants' differing interpretations of the task description (so that some participants focused only on *identifying* targets while others focused only on *locating* them), we analyze the distribution of strategies within subject. We examine the strategies chosen for all pairs consisting of a target and non-image-region landmark. All but 3 of 152 participants use more than one strategy, and the median number of strategies used is three (of the four total). This shows that subjects selected strategies in a scene- and target-dependent way, and thus variation does not reflect differences across participants in their interpretation of the task.

We conduct four one-vs.-all regression analyses to analyze which factors predict the choice of each order. The factors selected for analysis include measurements of visual salience (the area of the anchor and landmark bounding boxes, their distance to screen center (centr.) (calculated to the center of the object's bounding box), and a low-level salience score indicating pixel dissimilarity from the background. These properties are known to make objects more visually salient and easier to find (Wolfe, 2012), and to increase their chances of being chosen as landmarks (Kelleher et al., 2005; Golland et al., 2010; Clarke et al., 2013). We also include visual factors for the distance between the two objects, and for the signed left-right distance (in case the string ordering is affected by which object appears further left in the image). We also include the number of dependents (landmarks mentioned relative to the object in the description) as a linguistic factor. Large numbers of dependents tend to lead to a "heavier" phrase which is more likely to need its own clause, or to shift to the end of a sentence (White and Rajkumar, 2012). Finally, we include some task-based factors: whether the anchor is the overall target of the expression and whether the landmark is an object or an image region.

The low-level salience score used in this study is a computational measurement of how visually distinctive the object is, based on a comparison of its visual features with the rest of the image. The score used here differs from the Torralba et al. (2006) score used in Elsner et al. (2014), which was not found to be a significant predictor of ordering strategy. In this study, we compute an improved score by reanalyzing the Wally images with five low-level salience models, creating five salience maps for each image. The salience models used were: Achanta (Achanta et al., 2009), AIM (Bruce and Tsotsos, 2007), AWS (Garcia-Diaz et al., 2012), CovSal (Erdem and Erdem, 2013), RCS (Vikram et al., 2012), and SIG (Hou et al., 2012). Images were preprocessed by downsampling by a factor of four. For each salience map, we compute the mean salience within every labeled bounding box in the image. Since the output of the salience models is highly correlated, we then perform PCA (Principal Components

Analysis) on the scaled matrix of salience measurements and take the first principal component of the transformed data as a cross-model consensus salience score.

We transform area to square root area and log-transform distance (between objects) and centrality (distance from object to center of image) values. Centrality values are negated, so that higher numbers indicate more central objects. We then scale all continuous factors to zero mean and unit variance and deviation-code binary factors as –0.5, 0.5. We fit a binomial generalized linear model of the data, using uncorrelated random slopes and intercepts for speaker and item (Barr et al., 2013) using LME4 Bates et al. (2015)[3]. No interaction terms were included. Models for PRECEDE and FOLLOW converged using the default optimization settings. Models for PRECEDE-EST and INTER failed to converge with these settings. For these analyses, image regions were discarded from the dataset (since regions essentially always PRECEDE and never use these strategies); the coefficient for this effect is indicated as $-\infty$. Then the effects with the smallest coefficients were removed until convergence; these coefficients are shown as X. Significance of factor main effects was tested using ANOVA to compare a model including all factors and a model leaving out the factor of interest[4].

Results of the regression analysis appear in **Table 1**. The largest effects are those relating to image regions, which overwhelmingly occur in the PRECEDE order (15.68 PRECEDE vs. –16.42 FOLLOW). Area of the landmark also has a substantial effect; larger objects tend to PRECEDE (3.97) and INTERLEAVE (1.53) while smaller ones FOLLOW (−4.48). Objects with many dependents ("heavy" phrases) occur more often in PRECEDE-ESTABLISH constructions (1.31) and less often in INTERLEAVE and FOLLOW (–0.057, –0.075).

Smaller, but still significant, effects include anchor area; larger anchors are less likely to be PRECEDED by landmarks (–0.42) and more likely to be FOLLOWED (0.40). The target is more likely to INTERLEAVE around a landmark (0.84). Finally, the low-level salience score has slight effects for landmarks, but not for anchors: more visually distinctive landmarks are more likely to PRECEDE their anchors (0.22) and less likely to FOLLOW them.

No significant effect is found for either distance measurement.

3.3. Analysis

The strong effects of anchor and landmark area support the hypothesis that more visually salient objects are considered part of common ground and that speakers place them earlier in their descriptions. The effects of the low-level salience score, though weak, point in the same direction. The effects of centrality are counterintuitive (more central landmarks are less likely to PRECEDE). This pattern is difficult to explain, since increasing centrality normally makes objects more salient (Judd et al., 2012). We speculate that the effect might be due to the frequent use of region descriptors like "at the top right" to restrict attention to off-centered areas of the image.

While the low-level salience score has a significant effect, its contributions are minor. This may indicate that area, rather than overall visual salience, is indeed the major contributing factor for ordering. But this explanation fits poorly with both visual and linguistic theories, since it posits a special-case visual process and an exception to our usual understanding of how objects enter common ground. A better explanation is probably that computational salience modeling simply does not capture all the complex factors which make up visual distinctiveness in a domain like Where's Wally. Clarke and Keller (2014) show that many popular low-level salience models fail to account for viewer perceptions even in simple contrived stimuli. Thus, the composite score used in this analysis is likely capturing only some of the visual distinctiveness of objects in the scene.

The primary motivation for the PRECEDE-ESTABLISH construction appears to be linguistic; it occurs when the landmark itself has many dependent sub-landmarks and thus requires its own clause. It is less likely to be chosen if the anchor is large and easily spotted on its own (in which case the preferred order is FOLLOW). But it is also not as often selected for large landmarks (which don't require dependent sub-landmarks or their own clause). These findings are in accord with Ward and Birner (1995), who state that objects introduced by existential "there is" should be new to the discourse. The ESTABLISH strategy is a way of putting these important but hard-to-see landmarks on the left of the clause without marking them as common-ground information.

4. PERCEPTION STUDY

If speakers prefer to use the PRECEDE order for easier to find (larger and more salient) landmarks vs. the FOLLOW order for harder to find (smaller and less salient) ones, do these tendencies help listeners to find the target objects quickly? We conduct a visual search experiment using the Wally images and controlled linguistic stimuli to evaluate this hypothesis. Since area, centrality and low-level distinctiveness models gave equivocal results as proxies for visual salience in the previous section, in this experiment, we measure visual salience more directly. We use target-only and landmark-only visual search tasks as indicators of how easy each object is to see on its own, and analyze the relative descriptions in the context of these scores for their components.

4.1. Stimuli

Stimuli consist of a Where's Wally image paired with a referring expression. There are four conditions, illustrated with examples referring to **Figure 1**. We selected a single target and landmark in each image, so that the objects and attribute-based descriptions used in the TARGET and LANDMARK stimuli for a given scene also feature in the LANDMARK PRECEDES and LANDMARK FOLLOWS stimuli:

- TARGET: At the upper right, **the man holding the red vase with a stripe**.
- LANDMARK: At the upper right, *the sphinx*.
- LANDMARK PRECEDES: At the upper right, to the left of *the sphinx*, **the man holding the red vase with a stripe on it**.

[3] In LME4, the model is specified as $follow \sim area + (0 + area|speaker) + (0 + area|image) + \ldots + (1|speaker) + (1|image)$.

[4] *P*-values are presented without the Bonferroni correction for multiple comparisons. A set of 52 comparisons at the 0.05 level includes about three type II errors on average.

- LANDMARK FOLLOWS: At the upper right, **the man holding the red vase with a stripe** to the left of *the sphinx*.

The targets and landmarks are chosen to represent a range of relative size and perceived visual salience values, and to be approximately balanced across regions of the screen. In each case, the target person is one of the people used as targets in Clarke et al. (2013); when possible, the landmark is also one mentioned by speakers in the corpus, although in a few cases this was not possible since speakers did not mention a landmark of the desired size. Descriptions of targets and landmarks contained enough attributes to make them unambiguous in isolation (so that a relative description was an overspecification, not the only disambiguating detail).

All stimuli were read by a British English speaker. Recordings in the *landmark* condition are the fastest (mean length 2.6 s) followed by the *target* condition (3.0 s). The relative description cases are longer and therefore slower; when the landmark precedes, the mean length is 4.4 s while when it follows, the mean length is 4.2.

4.2. Experimental Procedures

The experiment was conducted in the Eye Movements and Attention laboratory at the University of Aberdeen. Experimental scripts were created and run using MatLab and run on a PowerMac. Stimuli were presented on a 61 cm Sony Trimaster EL computer screen, 1080 × 1920 computer screen. Participant responses were recorded using an Apple keyboard and mouse. An EyeLink 1000 was used to conduct eye-tracking, although eye-movements are not analyzed here. The protocol for each of the experiments was reviewed and approved by the Psychology Ethics Committee at the University of Aberdeen.

Thirty-two participants (median age 23, range = 19–42 years old, 21 females) took part in the study. Participants were recruited from the population of students and other members of the academic community at the University of Aberdeen. All participants had normal or corrected-to-normal vision and were native English speakers. The experiment was conducted with the full understanding and signed consent of each participant. Participants were remunerated $ 5–10 for their time, depending on the number of experiments they had taken part in.

Immediately following image onset, an audio recording of the search instruction was played to participants over headphones, giving them the necessary information required to find the target. Participants pressed the space bar on the keyboard when they had found the specified target. They were then required to use the mouse to click on the target. This was done so that we had a record of search accuracy and participants were not able to just press space without finding the target. Reaction time was recorded as the time from image onset to when the space bar had been pressed. There was no requirment for the participant to listen to the whole referring expression.

4.3. Outliers

The complete dataset consists of 896 trials (32 × 28). We filter the reaction time data from the perception study by discarding instances where the listener failed to find the target, or incorrectly signaled success before actually finding it. A single participant was discarded for excessively long reaction times. All trials for which the reaction time recorded was <0.5 s or >10 s were discarded, as were trials for which the time between the keypress signaling successful detection and the click to indicate the found item was greater than 5 s. These filters exclude 186 trials after which 669 remain. A software error prevented measurement of the click location for 56 trials, so we have accuracy information for only 613 of these.

4.4. Results

Overall, participants reacted faster to the non-relative expressions (median 3.9 s for targets and 3.7 for landmarks) than the relative ones (4.6 s for target-first REs and 4.9 for landmark-first REs). These times are approximately a second longer than the stimuli, and indicate that our visual search task was reasonably easy, especially given the cluttered nature of the scenes. In particular, the short search times for target-only expressions demonstrate that the relative descriptions were truly overspecified, since participants could find the targets without them. As usual in complex visual search tasks, standard deviations are substantial (between 1.0 and 1.3 for all cases).

Our analysis focuses on comparisons between the two orders for relative REs (PRECEDE and FOLLOW). We hypothesize that, when the target is easier to find than the landmark, search is facilitated by landmark FOLLOWING the target, while when the landmark is easier, search is facilitated by the landmark PRECEDING. We separate the stimuli into three categories, "target-easier," "target-harder," and "both-similar," based on the empirical reaction times for the target-only and landmark-only cases. For each image, we compute:

$$Z(median(rt_{targ-only}) - median(rt_{lmark-only})) \qquad (1)$$

This is a Z-transformed score of how much easier it is for participants to find the target than the landmark. We select the bottom third (nine instances) as "target-easier," the middle third (9 instances) as "both-similar," and the upper third (10 instances) as "target-harder."

Figure 3 shows a plot of reaction time as a function of referring expression order within each group. Median RTs are lower for the landmark FOLLOW order in the "both-similar" and "target-easier" groups and higher in the "target-harder" group. The overall median RT for the relative referring expressions is 4.7 s. In the "target-easier" group, the median for FOLLOW expressions is 4.3 while for PRECEDE expressions it is 4.9. For the "target-harder" group, the median for FOLLOW expressions is 5.3 while for PRECEDE expressions it is 4.7.

We perform the Mann-Whitney test for differing medians on each group. For the "both-similar" group, the test fails to find significance ($p > 0.05$); for the "target-easier" group, $p < 0.01$ and for the "target-harder" group, $p < 0.05$[5].

In addition to this analysis based on grouping the items, it is also possible to look at the median ($target - lmark$) (Equation 1) as a continuous predictor. In **Figure 4**, we plot it against the

[5] The null hypothesis for the "target-harder" medians cannot be rejected at a Bonferroni-corrected level of $0.05/3 = 0.016$.

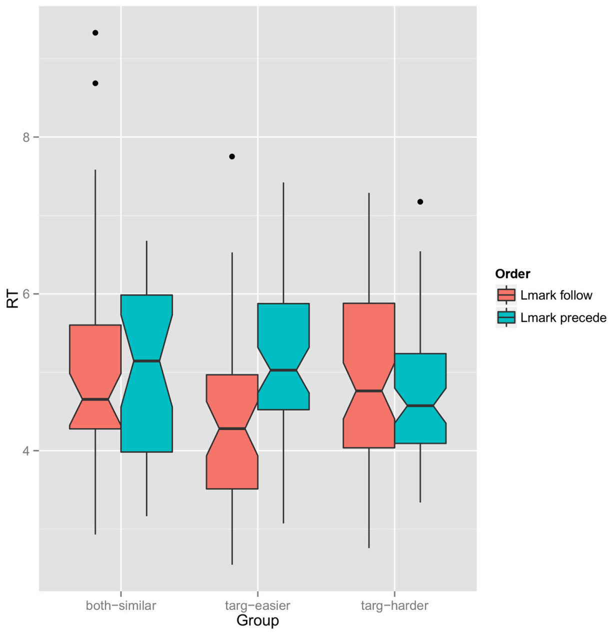


FIGURE 3 | Notched boxplot of reaction time as a function of referring expression order (red: target first, blue: landmark first) grouped by which object is easier to find. Notches represent 95% confidence interval of the median (computed with GGPlot default settings).

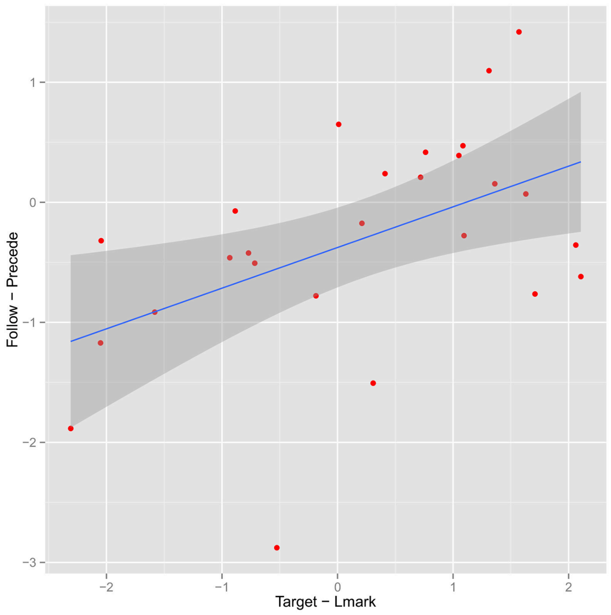


FIGURE 4 | Plot of median (target-first-landmark-first) reaction time as a function of median (target-landmark) reaction time. Each point represents a stimulus; fitted regression line uses linear model.

analogous quantity for the two relative referring expressions, median (*follow* − *precede*). Points on the left represent instances where the target is found faster than the landmark in isolation. Points at the bottom represent instances for which the FOLLOW order leads to a faster search. Thus, our hypothesis would predict a positive correlation. The estimated Pearson linear correlation is 0.52, (95% confidence interval 0.17–0.75).

Participants are relatively accurate (of 613 cases with accuracy information, 487 found the correct item with an error <150 pixels on either axis). We checked for an accuracy effect by group similar to the effect on reaction times, but there is none. Unsurprisingly, the majority of identification errors for relative descriptions (62 of 77) occur in the “target-harder” group, indicating that when the target takes longer to find, it is also more likely to be misidentified. But these are distributed evenly across the two RE orders[6].

4.5. Discussion

Under both analyses of the visual search study, the results are as predicted by our hypothesis: search is facilitated by mentioning the easier-to-find object first. The difference in medians suggests an average effect of about 0.6 s in either direction. Since the reaction time is measured from the start of the utterance, the results imply that giving the target description later in the trial can sometimes be beneficial, even though listeners in this condition must wait longer before they can possibly react.

Several caveats apply. First, although we find the expected facilitation effect when comparing among differently ordered relative descriptions, overall, participants reacted faster to the *non*-relative (target-only) expression. Even for the “targ-harder” group, mentioning the target alone yields a median search time of 4.2 s, while a relative description with the landmark first yields a median of 4.7.

If target-only descriptions actually lead to faster search than relative ones, why use a relative description at all? Clarke et al. (2013) show that relative descriptions are extremely common in human REs for these scenes, an effect also shown in a variety of previous work (Viethen and Dale, 2008). Overspecification is often intended to ensure the listener that they have actually found the right object (Arts et al., 2011; Koolen et al., 2011). If the listener believes confirmatory information is coming, they may wait to be sure they find the right object. However, Listeners are no more accurate in these conditions.

Secondly, the analysis does not correct for possible per-participant or per-item effects. This is partly due to the small amount of data, and partly to the use of median statistics to group the items as easier or harder. Since no participant heard more than one condition for a given stimulus, the easier/harder grouping reflects data from different participants than the reaction times plotted for relative descriptions within that group, complicating any analysis of individual differences.

5. CONCLUSION

Our analysis finds evidence for both of our hypotheses: speakers treat visually salient landmarks as being in common ground,

[6] We also ran the analyses above excluding trials on which a misidentification occurred; results are qualitatively similar, except that the test of whether median RTs differ in the “target-harder” group cannot be rejected.

preferring to place them early in their descriptions, and this ordering principle aids listeners in finding the target of a relative description quickly. These findings remain consistent with an audience-design model of perceptual effects in REG. In other words, speakers keep mental track of which objects in the scene are easier or harder to perceive. They use this information to preferentially select easier-to-see objects as landmarks, and they treat easier- and harder-to-see landmarks differently when planning the syntax of their descriptions. Both of these tendencies stem from the desire to make sure their listeners can efficiently find the object they are trying to point out.

While the results are consistent with such a model, we should emphasize that they do not rule out a least-effort model in which speakers talk more about things they themselves see earlier. To eliminate this possibility, we could give the speaker and listener different views of the scene [for instance, by occluding part of the scene for the listener (Brown-Schmidt et al., 2008)]. Alternately, we could look more closely at the time course of REG, using eye-tracking to determine when speakers discover the objects they mention and how much planning time intervenes.

Our findings definitely indicate that the choice of ordering strategy must be sensitive to visual features and cannot simply be left to an off-the-shelf micro-planning and realization component. This differentiates it from purely surface phenomena like dependency length minimization and heavy NP shift, which can be implemented at a late stage of the pipeline White and Rajkumar (2012). Choosing the correct strategy has a modest, but significant impact on listener performance. We find differences of about 0.6 s for referring expressions of about 4.7 s in length; in other words, the median subject's search will be about 10% easier if the correct ordering is used. Since we also found that relative descriptions lead to slower searches in general, this result should be considered with some caution. The stimuli used in this study were deliberately overspecified so that subjects could find the appropriate object using the non-relative description alone. Real relative descriptions are not always overspecified, but might be necessary to disambiguate the target; in these cases, they will presumably not cause a slowdown. The direction and magnitude of the slowdown effect might also vary depending on the complexity and visual clutter of the scene. Nonetheless, we believe that new REG systems should use perceptual information to properly order the relative descriptions they generate.

Our findings show that seemingly low-level and disparate mental "modules" like perception and sentence planning interact at a high level and in task-dependent ways. But we have yet to determine what sort of mental representations these systems use to communicate, or what underlies the considerable variation we find among both speakers and listeners. Our datasets are too small to tell us whether this variation reflects different populations, each using different strategies, or whether there is comparable variation within a single individual. Nor can it tell us whether larger-scale cognitive differences (for example, in attention, memory, or executive function) could account for these differences.

5.1. Data Sharing

The referring expressions used in the corpus study are publically available as the WREC (Wally Referring Expression Corpus): http://datashare.is.ed.ac.uk/handle/10283/337. See Clarke et al. (2013). The recorded stimuli used in the comprehension experiment are provided as Supplementary Material to this paper.

AUTHOR CONTRIBUTIONS

AC and ME designed the visual search comprehension study, which was run by student research assistants under the supervisor of AC. HR and ME selected and recorded the language stimuli. Analysis of the results was carried out by ME and AC. ME implemented the statistical analyses, while AC provided the visual salience information. The manuscript was jointly written by AC, ME, and HR.

ACKNOWLEDGMENTS

Thanks to Amelia Hunt for use of lab space. Warren James, Alex Irvine, and Melissa Spoliti helped with data collection. Thanks to Matt Stainer for helping with the implementation of the salience models and Laura Arnold for recording the comprehension stimuli. The open access publication of this work was generously supported by the College of Humanities and Social Sciences at the University of Edinburgh.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01793

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Süsstrunk, S. (2009). "Frequency-tuned salient region detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)* (Miami Beach, FL), 1597–1604. doi: 10.1109/CVPR.2009.5206596

Ariel, M. (1988). Referring and accessibility. *J. Linguist.* 24, 65–87. doi: 10.1017/S0022226700011567

Arts, A., Maes, A., Noordman, L., and Jansen, C. (2011). Overspecification facilitates object identification. *J. Pragmat.* 43, 361–374. doi: 10.1016/j.pragma.2010.07.013

Barclay, M. (2010). *Reference Object Choice in Spatial Language: Machine and Human Models.* PhD thesis, University of Exeter.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* 67, 1–48, doi: 10.18637/jss.v067.i01

Belz, A., and Gatt, A. (2007). "The attribute selection for GRE challenge: overview and evaluation results," in *Proceedings of UCNLG+ MT: Language Generation and Machine Translation* (Copenhagen), 75–83.

Belz, A., and Gatt, A. (2008). "Intrinsic vs. extrinsic evaluation measures for referring expression generation," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers* (Columbus, OH: Association for Computational Linguistics), 197–200.

Beun, R.-J., and Cremers, A. H. (1998). Object reference in a shared domain of conversation. *Pragmat. Cogn.* 6, 121–152. doi: 10.1075/pc.6.1-2.08beu

Brown-Schmidt, S., Gunlogson, C., and Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition* 107, 1122–1134. doi: 10.1016/j.cognition.2007.11.005

Bruce, N., and Tsotsos, J. (2007). Attention based on information maximization. *J. Vis.* 7, 950. doi: 10.1167/7.9.950

Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7

Clarke, A., Elsner, M., and Rohde, H. (2013). Where's Wally: the influence of visual salience on referring expression generation. *Front. Psychol.* 4:329. doi: 10.3389/fpsyg.2013.00329

Clarke, A. D. F., and Keller, F. (2014). "Measuring the salience of an object in a scene," in *Proceedings of Vision Science Society* (St. Pete Beach, FL).

Degen, J., and Franke, M. (2012). "Optimal reasoning about referential expressions," in *Proceedings of SEMDial* (Paris).

Duan, M., Elsner, M., and de Marneffe, M.-C. (2013). "Visual and linguistic predictors for the definiteness of referring expressions," in *Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)* (Amsterdam).

Duckham, M., Winter, S., and Robinson, M. (2010). Including landmarks in routing instructions. *J. Loc. Based Serv.* 4, 28–52. doi: 10.1080/17489721003785602

Elsner, M., Rohde, H., and Clarke, A. (2014). "Information structure prediction for visual-world referring expressions," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (Gothenburg: Association for Computational Linguistics), 520–529. doi: 10.3115/v1/E14-1055

Erdem, E., and Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *J. Vis.* 13, 11. doi: 10.1167/13.4.11

Frank, M. C., and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science* 336, 998–998. doi: 10.1126/science.1218633

Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., and Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: a computational approach. *J. Vis.* 12, 17. doi: 10.1167/12.6.17

Gkatzia, D., Rieser, V., Bartie, P., and Mackaness, W. (2015). "From the virtual to the realworld: referring to objects in real-world spatial scenes," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon: Association for Computational Linguistics), 1936–1942.

Golland, D., Liang, P., and Klein, D. (2010). "A game-theoretic approach to generating spatial descriptions," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (Cambridge, MA: Association for Computational Linguistics), 410–419.

Handford, M. (1987). *Where's Wally? 3rd Edn*. London: Walker Books.

Handford, M. (1988). *Where's Wally Now? 4th Edn.* London: Walker Books.

Handford, M. (1993). *Where's Wally? In Hollywood, 3rd Edn*. London: Walker Books.

Horton, W., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117. doi: 10.1016/0010-0277(96)81418-1

Horton, W. S., and Gerrig, R. J. (2002). Speakers experiences and audience design: knowing when and knowing how to adjust utterances to addressees. *J. Mem. Lang.* 47, 589–606. doi: 10.1016/S0749-596X(02)00019-0

Hou, X., Harel, J., and Koch, C. (2012). Image signature: highlighting sparse salient regions. *IEEE Trans. Anal. Mach. Intell.* 34, 194–201. doi: 10.1109/TPAMI.2011.146

Judd, T., Durand, F., and Torralba, A. (2012). *A Benchmark of Computational Models of Saliency to Predict Human Fixations.* Technical Report, MIT. Report no. MIT-CSAIL-TR-2012-001.

Kelleher, J., Costello, F., and van Genabith, J. (2005). Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artif. Intell.* 167, 62–102. doi: 10.1016/j.artint.2005.04.008

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Krahmer, E., and van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI_a_00088

Lockridge, C. B., and Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychon. Bull. Rev.* 9, 550–557. doi: 10.3758/BF03196312

Maienborn, C. (2001). On the position and interpretation of locative modifiers. *Nat. Lang. Semant.* 9, 191–240. doi: 10.1023/A:1012405607146

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics* 27, 3–1756.

Prince, E. (1981). "Toward a taxonomy of given-new information," in *Radical Pragmatics,* ed P. Cole (New York, NY: Academic Press), 223–255.

Roberts, C. (2003). Uniqueness in definite noun phrases. *Lang. Philos.* 26, 287–350. doi: 10.1023/A:1024157132393

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., and Kaufmann, S. (2012). "Communicating with cost-based implicature: a game-theoretic approach to ambiguity," in *The 16th Workshop on the Semantics and Pragmatics of Dialogue* (Paris), 107–116.

Talmy, L. (1978). "Figure and ground in complex sentences," in *Universals of human language*, Vol. 4, ed J. Greenberg (Stanford: Stanford University Press), 625–649.

Talmy, L. (1983). "How language structures space," in *Spatial Orientation: Theory, Research and Application,* eds H. L. Pick Jr and L. P. Acredolo (New York, NY: Springer), 225–282.

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. M. (2006). Contextual guidance of attention in natural scenes: the role of global features on object search. *Psychol. Rev.* 113, 766–786. doi: 10.1037/0033-295X.113.4.766

Viethen, J., and Dale, R. (2006). "Algorithms for generating referring expressions: do they do what people do?" in *Proceedings of the Fourth International Natural Language Generation Conference*, INLG '06 (Stroudsburg, PA: Association for Computational Linguistics), 63–70. doi: 10.3115/1706269.1706283

Viethen, J., and Dale, R. (2008). "The use of spatial relations in referring expressions," in *Proceedings of the 5th International Conference on Natural Language Generation* (Salt Fork, OH).

Viethen, J., and Dale, R. (2011). "GRE3D7: a corpus of distinguishing descriptions for objects in visual scenes," in *Proceedings of the Workshop on Using Corpora in Natural Language Generation and Evaluation* (Edinburgh: Association for Computational Linguistics).

Vikram, T. N., Tscherepanow, M., and Wrede, B. (2012). A saliency map based on sampling an image into random rectangular regions of interest. *Pattern Recognit.* 45, 3114–3124. doi: 10.1016/j.patcog.2012.02.009

Vogel, A., Potts, C., and Jurafsky, D. (2013). "Implicatures and nested beliefs in approximate decentralized-pomdps," in *ACL (2)* (Sofia: Citeseer), 74–80.

Ward, G., and Birner, B. (1995). Definiteness and the English existential. *Language* 71, 722–742. doi: 10.2307/415742

Ward, G., and Birner, B. (2001). "Discourse and information structure," in *Handbook of Discourse Analysis*, eds D. Schiffrin, D. Tannen, and H. Hamilton (Oxford: Basil Blackwell), 119–137.

White, M., and Rajkumar, R. (2012). "Minimal dependency length in realization ranking," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island: Association for Computational Linguistics), 244–255.

Wolfe, J. M. (2012). "Visual search," in *Cognitive Search: Evolution, Algorithms and the Brain*, eds P. Todd, T. Holls, and T. Robbins (Cambridge, MA: MIT Press), 159–175.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

HYPOTHESIS AND THEORY
published: 07 April 2016
doi: 10.3389/fpsyg.2016.00506

Cognitive Modeling of Individual Variation in Reference Production and Comprehension

*Petra Hendriks**

Center for Language and Cognition Groningen (CLCG), University of Groningen, Groningen, Netherlands

A challenge for most theoretical and computational accounts of linguistic reference is the observation that language users vary considerably in their referential choices. Part of the variation observed among and within language users and across tasks may be explained from variation in the cognitive resources available to speakers and listeners. This paper presents a computational model of reference production and comprehension developed within the cognitive architecture ACT-R. Through simulations with this ACT-R model, it is investigated how cognitive constraints interact with linguistic constraints and features of the linguistic discourse in speakers' production and listeners' comprehension of referring expressions in specific tasks, and how this interaction may give rise to variation in referential choice. The ACT-R model of reference explains and predicts variation among language users in their referential choices as a result of individual and task-related differences in processing speed and working memory capacity. Because of limitations in their cognitive capacities, speakers sometimes underspecify or overspecify their referring expressions, and listeners sometimes choose incorrect referents or are overly liberal in their interpretation of referring expressions.

Keywords: ACT-R, cognitive modeling, perspective taking, processing speed, reference comprehension, reference production, working memory

OPEN ACCESS

Edited by:
Kees Van Deemter,
University of Aberdeen, UK

Reviewed by:
Agnieszka Konopka,
Max Plank Institute
for Psycholinguistics, Netherlands
Antal Van Den Bosch,
Radboud University Nijmegen,
Netherlands

****Correspondence:***
Petra Hendriks
p.hendriks@rug.nl

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

Received: *31 July 2015*
Accepted: *24 March 2016*
Published: *07 April 2016*

Citation:
Hendriks P (2016) Cognitive Modeling
of Individual Variation in Reference
Production and Comprehension.
Front. Psychol. 7:506.
doi: 10.3389/fpsyg.2016.00506

LINGUISTIC REFERENCE

An important function of language is reference. Speakers refer to things, people, or events in the world around them and listeners identify these referents based on the referring expressions used by the speaker. To refer, speakers can use a variety of forms. For example, to refer to their neighbor they could utter the indefinite noun phrase *a lady who lives next door* or the definite noun phrase *the lady who lives next door*, refer to her by her name, use a personal pronoun such as *she* or *her* or a reflexive pronoun such as *herself*. Which form a speaker decides to use depends on a large number of factors, including the structure of the sentence and the prominence of the referent in the context of utterance. Likewise, to interpret the referring expression uttered by the speaker, listeners can often choose between various referents. This choice also depends on various factors.

Reference has been a central topic in many subfields of linguistics over the past decades. Although the factors influencing speakers' production and listeners' comprehension of referring expressions have been studied extensively from various angles, there does not exist a comprehensive account of linguistic reference yet. One of the main challenges for such a comprehensive account is the observation that speakers vary considerably in their choice of referring expression. This variation is problematic for most theoretical and computational models of linguistic reference. Theoretical models of referential choice (e.g., Gundel et al., 1993, 2012) generally attribute this variation to the interaction of the model with general cognitive and pragmatic factors and principles, without offering a specification of how the variation arises. Most computational algorithms for the generation of referring expressions are deterministic (van Deemter et al., 2012) and therefore always generate the same referring expression in a particular situation (but see Frank and Goodman, 2012; van Gompel et al., 2012; Mitchell et al., 2013, for recent probabilistic approaches). As a consequence, no variation is produced. Even if different individual speakers are modeled by different computational algorithms (as proposed by Dale and Viethen, 2010), variation within speakers is not accounted for.

This paper addresses the question of how the observed variation in speakers' choice of referring expression and listeners' choice of referent (which is discussed below) can be accounted for. It is hypothesized that at least part of this variation can be explained as resulting from the dynamic interaction between linguistic and cognitive constraints on reference. The application of these constraints is dependent on the cognitive capacities of speakers and listeners. Cognitive capacities can vary between individuals (e.g., between children and adults), but also within the same individual (e.g., due to linguistic and cognitive development or as an effect of the task). For example, some tasks are cognitively more demanding than other tasks and will therefore leave the speaker or listener with insufficient cognitive resources to make an optimal referential choice. A promising approach to investigate this hypothesis is by computational cognitive modeling of linguistic reference. Using computational cognitive modeling, models are developed that are cognitively plausible rather than computationally optimal. Hence, they incorporate the normal variability found in human performance.

In the next section, we discuss different types of variation that have been observed in the psycholinguistic literature on referential choice. Next, the cognitive architecture ACT-R is introduced. Within this cognitive architecture, a computational cognitive model of reference has been developed. It is shown how this ACT-R model is able to account for speakers' underspecification and overspecification of referring expressions and listeners' incorrect interpretations of referring expressions in particular tasks.

VARIATION IN REFERENTIAL CHOICE

Psycholinguistic investigations of the referential choices made by human speakers reveal considerable variation, both among and within individuals and across tasks. For example, in a web-based experiment where adult participants were asked to produce referring expressions to describe one of three objects shown in a picture in such a way that a friend looking at the same picture would be able to identify the target referent, speakers were found to show a large amount of variation in the referring expressions they produced (Dale and Viethen, 2010). For the same visual scene, different speakers produced different forms, such as *the blue cube*, *the blue cube in front of the red ball*, and *the cube in front of the ball*. Many of these forms were overspecific and informationally redundant. A few were not specific enough and failed to uniquely distinguish the target referent from the other two objects. This illustrates that, even in the very same task, speakers overspecify as well as underspecify their referring expressions. Similar patterns of frequent overspecification and some underspecification of referring expressions were found in other psycholinguistic studies (e.g., Deutsch and Pechmann, 1982; Engelhardt et al., 2006; Koolen et al., 2011).

In addition to variation among speakers, it has been observed that there is also variation within speakers. van Deemter et al. (2012, p. 174) examined the data of Fukumura and van Gompel (2010), who carried out two written sentence completion experiments with adult participants to investigate the choice between a pronoun and a name for a previously mentioned referent. Van Deemter and colleagues found that the majority of participants in the study of Fukumura and van Gompel did not produce only pronouns or only names, but produced both types of referring expression in at least one of the conditions of the experiments.

Variation in speakers' referential choices can also be observed in more naturalistic tasks, such as telling a story. Reference production during story telling is often investigated on the basis of cartoon movies or picture books (e.g., Karmiloff-Smith, 1985; Arnold et al., 2009). For example, picture stories like **Figure 1** were used to elicit narratives in children, young adults and elderly adults (Hendriks et al., 2014), and in children with autism, children with ADHD and typically developing children (Kuijper et al., 2015). The picture stories in these experiments featured two characters of the same gender and were designed to elicit two topic shifts: one halfway through the story (when the second character enters the story) and the other one at the end of the story (when there is a shift in focus to the first character again). The participants were instructed to tell the story so that a second experimenter, who could not see the pictures, would be able to understand the story.

On the basis of the picture story in **Figure 1**, participants produced narratives such as the following (from Hendriks et al., 2014, translated from Dutch and slightly adapted for the sake of readability):

> Speaker 1: A pirate with the football. Then he kicks it. Then it is in the water. Then the knight goes to catch it. And he has caught the ball in a net. Now he has his ball back again.
>
> Speaker 2: The pirate with a wooden leg has a football. He kicks the football with his wooden leg into the pond. And cries because he can't reach the ball anymore. The knight sees all that. The knight gets a net. And gets the ball out of the water for the pirate. The pirate has a big smile because he is happy that he has the ball back again.

The narrative produced by speaker 2 differs from the narrative produced by speaker 1 in several respects: Speaker 2 uses longer sentences than speaker 1 with more variation in their structure, speaker 2 provides more information than speaker 1 and explicitly mentions causal relations between events (as, e.g., marked by the causal connective *because*), and speaker 2

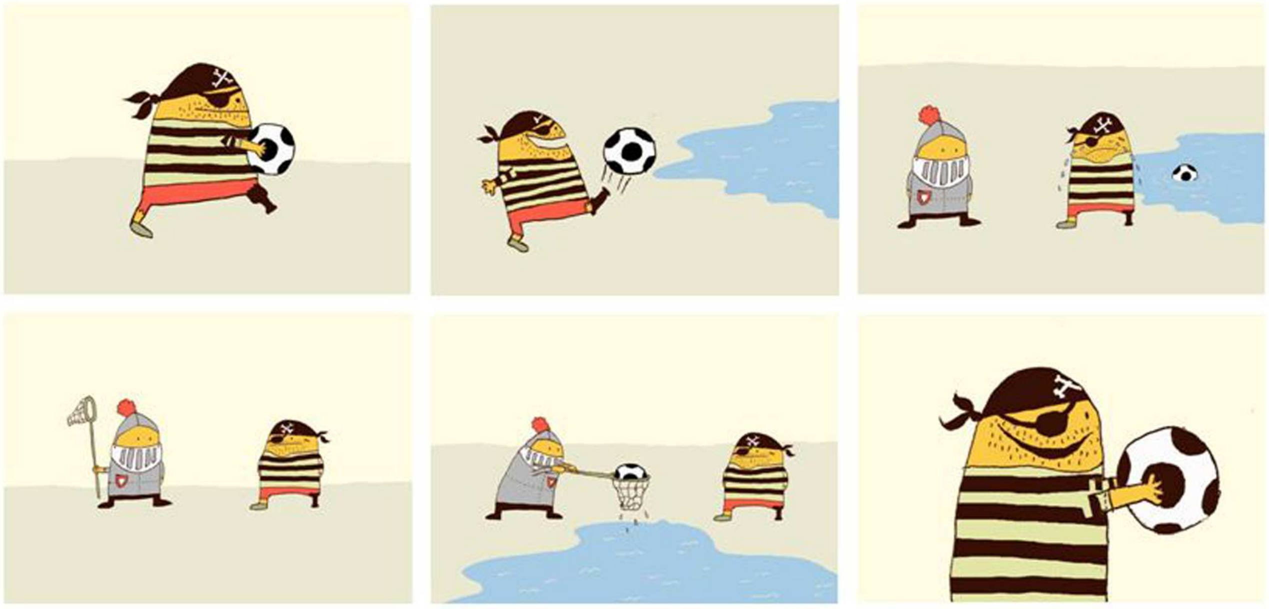


FIGURE 1 | Picture story used for eliciting narratives in the studies of Hendriks et al. (2014) and Kuijper et al. (2015).

makes different referential choices than speaker 1 and uses fewer pronouns and more full noun phrases.

One reason for the observed differences between the narratives of the two speakers is that speaker 1 is a 6-year-old child, while speaker 2 is a 27-year-old adult. As an adult can be expected to have more linguistic experience than a child speaker and also is expected to possess more cognitive resources, this may be the reason for the longer and more elaborate sentences and the more explicit references in the narrative produced by speaker 2. However, suggesting a potential reason is not the same as providing an explanation. In particular, the observation that the two speakers differ in age does not tell us which aspects of children's linguistic or cognitive abilities must develop further to result in more adult-like narrative productions.

In addition to variation among speakers, the two narratives also illustrate another type of variation, namely variation across tasks. When telling a story, speakers must introduce the characters in the story as referents in the linguistic discourse, maintain reference to these characters when talking about their actions, and occasionally shift the attention of the listener from one character to the other. These actions can be considered as separate tasks carried out by the same speaker. Crucially, these tasks are subject to different constraints. To introduce the two characters in the story, both speakers use a full noun phrase. On the other hand, to continue to refer to the two characters, the adult speaker uses pronouns as well as full noun phrases, whereas the child speaker only uses pronouns. As the referents of these pronouns may not always be uniquely identifiable for a listener, these pronouns may underspecify the intended meaning. So the adult speaker and the child speaker make similar referential choices on the task of introducing referents, but make different referential choices on the task of maintaining and shifting reference. This illustrates that there is an intricate interaction between a speaker's linguistic and cognitive abilities and the properties of the task.

To investigate the interaction between linguistic constraints, cognitive constraints and task effects on referential choice, it is useful to combine psycholinguistic experimentation with computational modeling. Using computational modeling can help in teasing apart the different factors involved in referential choice and shed more light on the way they interact. This may contribute to our understanding of why speakers overspecify and underspecify their referring expressions and why different speakers do so to different degrees. Also, computational modeling may reveal how the speaker's referential choices affect the listener. The paper will focus on the type of referring expression and *how* speakers refer (e.g., with a full noun phrase or a pronoun) and how listeners interpret such referring expressions. Although the speaker's choice of *what* to refer to (e.g., whether to express a causal relation between two events or not, or whether to refer to the pirate or the knight) also is an important aspect of referential choice, this is beyond the scope of this paper.

COMPUTATIONAL MODELING IN ACT-R

Computational modeling of language often has the practical goal of developing computational algorithms that can be used in natural language applications. Virtually all natural language generation systems contain a module for generating referring expressions (Mellish et al., 2006). Although the aim of these systems is to be practically useful, computational models for the generation of referring expressions are usually not evaluated in terms of their usefulness, but instead in terms of their human-likeness (see Krahmer and van Deemter, 2012, for a comprehensive survey). In particular, these computational models aim to mimic human performance such as reflected in the group results of behavioral studies or in the patterns found in a corpus of written texts. For example, Kibrik and colleagues (Kibrik et al., 2013) developed a computational model that, using classical machine learning algorithms, determines referential choice in discourse on the basis of multiple factors related to properties of the referent and the discourse context. As these factors and their weights are extracted from a corpus of newspaper articles, they pertain to general patterns of referential choice generated by multiple writers on multiple occasions, as opposed to the specific referential choices made by individual writers in particular situations. However, if speakers and writers show variation in their referential choices, the specific patterns produced by individual speakers or writers may differ from the general patterns observed at the group level. Furthermore, computational models of this type are usually evaluated on the basis of their similarity with human offline referential choice only, rather than on the basis of human online referential processing as well. Also, they are generally concerned with either language production or language comprehension, but not both.

In addition to its use in natural language applications, computational modeling of language can also be useful for the development of psycholinguistic theories. Regarding reference, the aim of such computational models is to mimic human offline as well as online referential processes. Computational modeling of language for psycholinguistic research makes it possible to assess the completeness of a theoretical account, forces the modeler to be precise, allows for the systematic manipulation of factors, and makes possible the generation of novel predictions. In this paper, we discuss a series of computational models of reference production and comprehension that have been implemented in the cognitive architecture ACT-R (Adaptive Control of Thought-Rational; Anderson et al., 2004; Anderson, 2007). Computational modeling in ACT-R has the additional advantage that ACT-R not only is a computational modeling environment, but also is a theory of human cognition in which detailed assumptions about cognitive processes have been implemented that are based on a range of data from psychological and neurocognitive experiments. As a consequence, ACT-R's modeling environment constrains the computational models in such a way that the models are cognitively plausible and are consistent with what is currently known about human cognition.

ACT-R is a hybrid architecture that combines symbolic and subsymbolic structures and processes. Whereas the chunks of factual information and the if-then production rules of ACT-R are symbolic in nature, certain processes of ACT-R are subsymbolic. When more than one production rule can be applied, there will be competition among these rules. The production rule with the highest expected utility will be executed. This is a subsymbolic process that is computed on the basis of mathematical equations weighing the costs of executing the production rule against its benefits. Another process that is dependent on properties at the subsymbolic level is the retrieval of chunks from declarative memory. Whether and how fast a chunk is retrieved depends on its activation value, which is a function of its frequency, its recency of use, and its connections to other chunks in memory.

A fundamental property of ACT-R is the assumption that each operation of the model takes time to perform. Every retrieval of a fact from declarative memory and every execution of a production rule takes a certain amount of time. Hence, performance of the model is limited by the time available for the cognitive process. However, the total execution time of the cognitive process is not simply the sum of the durations of all constituting operations. This is because the different modules of ACT-R can operate in parallel, although each module by itself can only perform a single operation at a time. Thus, the duration of a cognitive process critically depends both on the timing of the serial processes within a module and on how the different modules interact. Furthermore, there is some random variation in the model, as the utilities associated with production rules and the activation values of chunks are noisy. Therefore, to provide specific time estimations for a cognitive process, simulations should be run with the computational model (Anderson et al., 2004).

An ACT-R model obtains higher processing efficiency and performs faster by means of the ACT-R learning mechanism of production compilation (Taatgen and Anderson, 2002). In production compilation, two existing production rules are integrated into one new production rule. Because fewer production rules are needed with this new single production rule than with the old two production rules, the result is faster and more automatic processing. Production compilation occurs when two existing production rules are repeatedly executed in sequence. Ultimately, as a result of production compilation, carrying out a cognitive task may not require retrieval of individual chunks from memory or execution of multiple production rules anymore, but may be done by a single general production rule.

The predictions of an ACT-R model can be tested by comparing the results of computational simulations of the ACT-R model on a specific cognitive task with the results of human participants carrying out the same task. The output of a simulation in ACT-R consists of quantitative measures of performance on the task and estimates of the time it takes to perform the task. Each simulation of the model simulates the performance of an individual participant on a task. By offering different amounts of training, the model can also simulate the performance of individual children of different ages (van Rij et al., 2010). Due to the random variation present in the model, performance of the model differs slightly during each run. Thus, ACT-R models are non-deterministic. By running an ACT-R model several times on the experimental items of a linguistic task, a dataset is obtained that can be compared to – and analyzed in the same way as – the dataset obtained from a group of human participants on the same task.

Because of the cognitive constraints placed on computational models in ACT-R, ACT-R can shed more light on the cognitive processes involved in language and communication (cf. Taatgen and Anderson, 2002; Budiu and Anderson, 2004; Lewis and

Vasishth, 2005; Reitter et al., 2011; Guhe, 2012). In particular, ACT-R's assumptions regarding the duration of cognitive operations allow us to make precise predictions about the time course of language processing. Furthermore, ACT-R makes it possible to integrate linguistic analyses of referential choice in the model, implement the opposite processes of language production and language comprehension in one and the same model, and describe the development and processing of perspective taking in language without additional assumptions. Cognitive modeling in ACT-R may therefore reveal the mechanisms underlying the observed variation in speakers' and listeners' referential choices.

COGNITIVE MODELING OF REFERENCE PRODUCTION AND COMPREHENSION

In a series of studies (Hendriks et al., 2007; van Rij et al., 2010, 2013; van Rij, 2012; Vogelzang et al., 2015), computational models have been implemented within the cognitive architecture ACT-R to simulate the production and comprehension of referring expressions by adults and children. These computational simulations focused on the type of referring expression (definite noun phrase, overt pronoun or null pronoun) in production and on the identification of the referent in comprehension. The outcomes of these simulations were compared to existing data and further simulations were run to generate new predictions. The details of the various ACT-R models of reference are presented below. As these models are based on the same principles, we will refer to them as the ACT-R model of reference, only mentioning differences between the models when relevant. Following the presentation of the ACT-R model of reference, it is discussed how the model explains individual variation in reference production and comprehension and how the model generates novel predictions that can be tested empirically.

Performance of the ACT-R model of reference proceeds in three steps (see **Figure 2**): (1) determining the topic of the linguistic discourse on the basis of general memory principles, (2) applying the linguistic constraints that underlie the choice and interpretation of referring expressions, and (3) considering the opposite perspective in communication, which provides internal feedback to the model on the correctness of the referential choice.

These three steps are discussed in more detail below. Particular emphasis is placed on the cognitive principles and mechanisms that are implemented in the model and that may be relevant for reference production and comprehension.

Step 1: Determining the Current Discourse Topic

The first step of the ACT-R model of reference (see, e.g., van Rij et al., 2013) consists of determining the current discourse topic. Using the general memory principles of ACT-R, the model incrementally builds a (simplified) representation of the linguistic discourse during online processing. Each discourse referent that is encountered is represented as a chunk in declarative memory that has a certain amount of activation. Within ACT-R, the activation of a chunk depends on its frequency of use and the recency of the last retrieval of the chunk. The more frequently the chunk is used, or the more recent its last retrieval, the higher its activation. The activation of a chunk decays with time, but is increased when the chunk is retrieved again. The discourse referent with the highest level of activation in declarative memory is taken to be the current discourse topic. This allows the model to use gradient information about the activation of referents for making discrete decisions about the linguistic effects of discourse topicality in the next step of the model.

In addition to this mechanism of *base-level activation*, ACT-R also has a mechanism of *spreading activation*, that can temporarily increase the activation of a chunk. Spreading activation reflects the usefulness of a chunk in a particular context: chunks that are currently being processed spread activation to connected chunks in declarative memory. In van Rij et al.'s (2013) model, the subject of the previous sentence is temporarily stored as goal-relevant information and therefore spreads activation to connected chunks. This reflects the observation that the subject of the previous sentence is likely to be the current discourse topic (e.g., Grosz et al., 1995). Because the referent that was mentioned as the subject of the previous sentence becomes more activated in comparison to other referents due to spreading activation, the model will more often select this referent as the discourse topic.

Building a representation of the discourse requires access to memory resources, which can be different for different individuals. ACT-R does not have a separate working memory (WM) component. However, one of the ways to model WM effects in ACT-R is through individual differences in spreading activation (van Rij et al., 2013). The amount of spreading

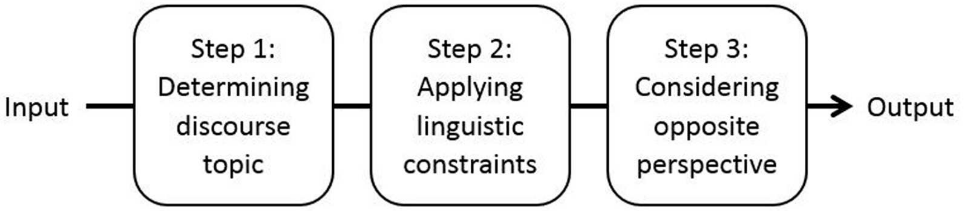

Input
Step 1: Determining discourse topic
Step 2: Applying linguistic constraints
Step 3: Considering opposite perspective
Output

FIGURE 2 | Performance of the ACT-R model of reference. Performance of the model proceeds in three steps. In production, the input is a meaning and the output is the optimal form for expressing this meaning. In comprehension, the input is a form and the output is the optimal meaning assigned to this form.

activation determines the ability to maintain goal-relevant information, and differences in the total amount of spreading activation account for individual differences in WM capacity (Daily et al., 2001). Hence, the effects of WM capacity on discourse processing can be modeled as resulting from differences in the ability to maintain goal-relevant information pertaining to the subject of the previous sentence (van Rij et al., 2013). In the ACT-R model of reference, a high WM capacity gives rise to a large amount of spreading activation of the chunk representing the subject of the previous sentence. This results in this previous subject being a determining factor in the selection of the discourse topic. In contrast, a low WM capacity only gives rise to a small amount of activation, resulting in no effect at all of the subject of the previous sentence on the selection of the discourse topic. In the latter case, frequency and recency will be the main determinants of the discourse topic.

The mechanism of base-level activation in combination with spreading activation implements the effects of the preceding linguistic discourse on the prominence, or accessibility, of discourse referents. Referents are more accessible if they are more frequently referred to, more recently referred to, or mentioned as the subject of the preceding sentence (cf. Givón, 1983; Ariel, 1988, 1990; Grosz et al., 1995; Arnold, 2010). Furthermore, influences of WM capacity on the selection of the discourse topic are predicted.

Step 2: Applying Linguistic Constraints on Referential Choice

The second step of the ACT-R model of reference consists of the application of linguistic constraints that restrict the choice and interpretation of referring expressions. These linguistic constraints and the way they interact are taken from Optimality Theory (Prince and Smolensky, 2004) and from theoretical analyses of referential choice in this linguistic framework. Constraints in Optimality Theory differ from rules in rule-based linguistic frameworks in that these constraints are formulated as general as possible and hence can be in conflict. Crucially, the constraints differ in strength and are violable. If two constraints are in conflict and cannot be satisfied both, the stronger constraint is satisfied at the cost of violating the weaker constraint. A second difference between linguistic constraints and linguistic rules is that, whereas linguistic rules are input-oriented, linguistic constraints are output-oriented. Rules apply if the input conditions are met. Constraints, on the other hand, apply if the output has particular features. For example, a constraint prohibiting the use of pronouns will apply if a potential output contains a pronoun. This property of constraints allows Optimality Theory to explain mismatches – that is, asymmetries – between production and comprehension in child language (Smolensky, 1996; Hendriks, 2014), as is explained below.

To produce or interpret a referring expression, the ACT-R model evaluates potential outputs for a particular input. In production, the input meaning is given and potential forms for expressing this meaning compete. On the basis of the constraints of the grammar, the optimal form for expressing the input meaning is selected from a set of competing forms. The optimal form is the form that satisfies the constraints of the grammar best. Whereas in production the input consists of a meaning and the output is the optimal form for this meaning, in comprehension the input consists of the form to be interpreted and the output is the optimal meaning for this form. Determining the optimal meaning in comprehension is subject to the same hierarchy of constraints as in production. Thus, production and comprehension are guided by the same grammar and only differ in the direction of optimization (from input meaning to optimal form versus from input form to optimal meaning).

In the ACT-R model, candidate forms, candidate meanings and linguistic constraints are implemented as chunks in declarative memory (see Misker and Anderson, 2003, for an alternative approach to combining ACT-R with Optimality Theory). Rather than determining the optimal candidate by simultaneously comparing all candidate outputs with respect to the complete hierarchy of constraints, as is assumed in theoretical work in Optimality Theory (Prince and Smolensky, 2004), the ACT-R model compares only two candidates at a time, starting with the candidates with the highest activation. Each of these two candidates is evaluated on the basis of only one constraint at a time, starting with the strongest constraint. If one of the two candidates satisfies this constraint and the other does not, this other candidate is discarded and a new candidate is retrieved from memory. If the two candidates both violate or satisfy the constraint, a next constraint is retrieved. The two candidates are then evaluated on the basis of this next constraint. By iteratively applying this procedure (see **Figure 3**), given sufficient time all candidates can be evaluated with respect to all constraints. The optimization procedure terminates if an optimal candidate has been found or if time is up. In the latter case, one of the two candidates under consideration is selected at random.

Two linguistic constraints that have been argued to be relevant for the production and comprehension of referring expressions in discourse (e.g., Hendriks et al., 2008) are Referential Economy (referentially less informative forms such as pronouns are preferred to referentially more informative forms such as full noun phrases) and ProTop (pronouns refer to the discourse topic). The former constraint is theoretically modeled in Optimality Theory as a family of constraints of differing strengths prohibiting referring expressions. As the constraint prohibiting full noun phrases is stronger than the constraint prohibiting pronouns, it is better to use a pronoun than to use a full noun phrase.

On the basis of these two constraints, an overall preference is predicted for producing pronouns, even for referents that are not highly prominent in the discourse. Furthermore, it is predicted that all pronouns are interpreted as referring to the discourse topic, that is, the most prominent referent in the discourse. This asymmetric pattern in the production and comprehension of anaphoric pronouns is consistent with the literature on children's use and interpretation of pronouns in discourse (e.g., Karmiloff-Smith, 1985; Song and Fisher, 2005). For example, Karmiloff-Smith (1985) notes that, in narrative production, 4-year-olds produce strings of pronouns that at times refer to the main character of the story and at other times to the subsidiary

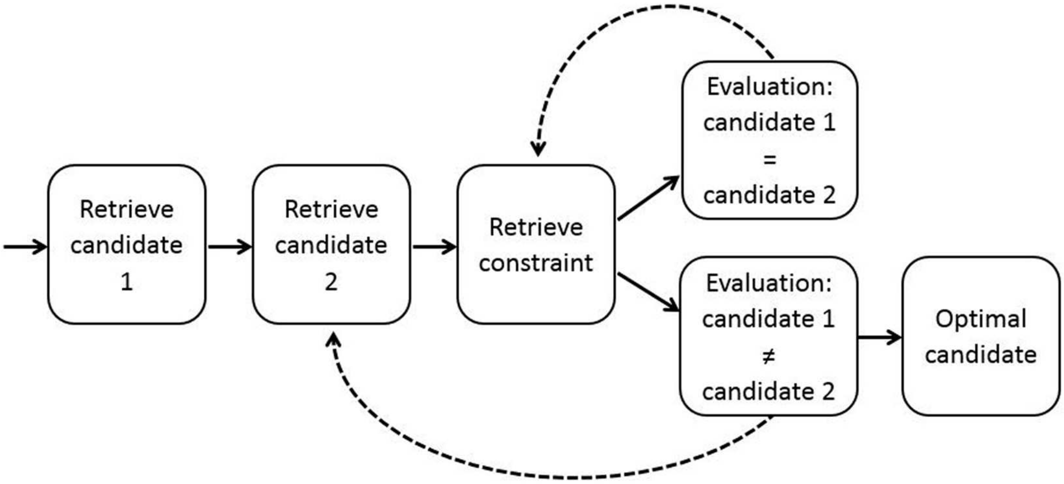


FIGURE 3 | Selection of the optimal candidate in the ACT-R model of reference. This process of optimization occurs in Steps 2 and 3. After retrieval of two candidates and a constraint, the two candidates are evaluated on the basis of the constraint. This procedure is applied iteratively, retrieving new candidates and new constraints until the optimal candidate is found or time is up. Adapted from Hendriks et al. (2007).

character, thus making reference ambiguous for the listener. On the other hand, 3-year-olds' comprehension of pronouns already depends in an adult-like way on the prominence of the referents in the linguistic discourse (Song and Fisher, 2005). So, in child language correct production of anaphoric pronouns seems to lag behind their correct comprehension. This particular asymmetry between production and comprehension is predicted by the two constraints mentioned above.

In contrast to children, adults do not show an overall preference for producing pronouns, regardless of the discourse context. Instead, their referential choices in production match their referential choices in comprehension. This is the motivation for the third step of the model, which further restricts adults' production of anaphoric pronouns by means of perspective taking.

For the production and comprehension of pronouns in syntactic binding environments, such as *her* in the sentence "*Goldilocks washed her*," the additional syntactic constraint Principle A is relevant. This constraint requires a reflexive to be bound within its clause (cf. Chomsky, 1981). That is, it requires *herself* in the sentence "*Goldilocks washed herself*" to be coreferential with the local subject *Goldilocks*. As Principle A is stronger than the constraint from the constraint hierarchy Referential Economy that prefers reflexives to pronouns, the two constraints together predict that local binding is expressed by reflexives and that reflexives are interpreted as being locally bound. Furthermore, these constraints predict that unbound referents are expressed by pronouns, as Principle A does not allow reflexives to appear unbound. This pattern is indeed observed in English-speaking children and adults. However, English-speaking children differ from adults in their interpretation of pronouns in syntactic binding environments. For children, such pronouns are ambiguous and can receive a bound as well as an unbound interpretation. That is, they take *her* in "*Goldilocks washed her*" to be able to refer to Goldilocks too. This asymmetric pattern is again predicted by the constraints and is generally known as the Delay of Principle B Effect, referring to the delayed development of object pronouns compared to reflexives in languages such as English and Dutch (Chien and Wexler, 1990; van Rij et al., 2010). In contrast to the asymmetry with anaphoric pronouns discussed above, in case of object pronouns in syntactic binding environments correct comprehension surprisingly lags behind correct production (De Villiers et al., 2006; Spenader et al., 2009). Because the interpretation of object pronouns is not restricted by syntactic constraints, object pronouns are allowed to be coreferential with the local subject. However, for adults this is not true. Again, the third step of the model is needed to further restrict adults' interpretation of pronouns.

The second step of the ACT-R model of reference crucially relies on linguistic knowledge. This implies that cross-linguistic differences in referential choice must receive their explanation in this part of the model. For example, the fact that in languages such as English sentences must always have a subject, whereas in languages such as Italian pronominal subjects can be dropped, can be explained by a different ranking of the same two constraints (Grimshaw and Samek-Lodovici, 1998). The availability of the additional possibility for expressing the subject in Italian as a null pronoun not only influences the distribution of overt pronouns, but may also influence the way these overt pronouns are interpreted (Vogelzang et al., 2015). As the three steps of the model are closely connected, these cross-linguistic differences in the constraint hierarchy and the inventory of linguistic forms are expected to also affect the other steps of the model.

Step 3: Considering the Opposite Conversational Perspective

The third step of the ACT-R model of reference is the consideration of the opposite perspective in communication.

After an initial choice has been made by the model in Step 2, the opposite communicative perspective is taken to verify whether this initial choice is also optimal from the opposite perspective. In production, the model first takes the perspective of the speaker to select a referring expression, and next takes the opposite perspective of a listener to check whether this referring expression is understandable for a hypothetical listener in the (speaker's representation of the) current linguistic discourse. Likewise, in interpretation, the model first takes the perspective of the listener to select a referent for the referring expression that is encountered, and next takes the opposite perspective of a speaker to check whether a hypothetical speaker would indeed have chosen this expression to refer to the selected referent in the (listener's representation of the) current linguistic discourse.

This mechanism of perspective taking is a serial implementation of the algorithm of bidirectional optimization in Optimality Theory (Blutner, 2000). Bidirectional optimization considers all pairs of linguistic form and meaning simultaneously and identifies the optimal pairs. It has the effect that if a form or meaning already is part of an optimal form-meaning pair, its use is blocked for another form-meaning pair. In the ACT-R model, bidirectional optimization is implemented as a serial process of perspective taking, starting with optimization from the language user's own perspective followed by optimization from the opposite perspective (Hendriks et al., 2007). This two-step process of perspective taking proceeds incrementally. That is, perspective taking is not postponed until the end of the sentence and also does not consider all pairs of form and meaning in one step (as in Blutner's bidirectional optimization algorithm), but rather is applied online and only considers two possibilities at a time. The extra step of optimization from the opposite perspective proceeds in the same way as optimization from the own perspective (as shown in **Figure 3**), after which the output of the extra step of optimization is compared to the input of the initial step of optimization (see **Figure 4**). If the output of the extra step of optimization (Step 3) differs from the input of the initial step of optimization (Step 2), the initially selected form or meaning is discarded and the next best form or meaning is taken as the input to the extra step of optimization. This process is repeated iteratively until output and input match or until time is up.

The mechanism of perspective taking thus generates internal feedback for the model. This feedback takes the form of a match or mismatch between the form-meaning pair resulting from optimization from one's own perspective and the form-meaning pair resulting from optimization from the opposite conversational perspective. A match results in an update of the parameters associated with the production rules that were used, increasing the chances that these production rules are used again next time. Mismatches have the effect that forms whose meaning is not recoverable for a listener and interpretations that are not expressed with the heard form by a speaker are blocked.

Obviously, the extra step of perspective taking takes additional time. Therefore, performing both steps (the step of the initial selection of the form or meaning and the additional step of perspective taking) during online production and comprehension requires sufficient processing speed (van Rij et al.,

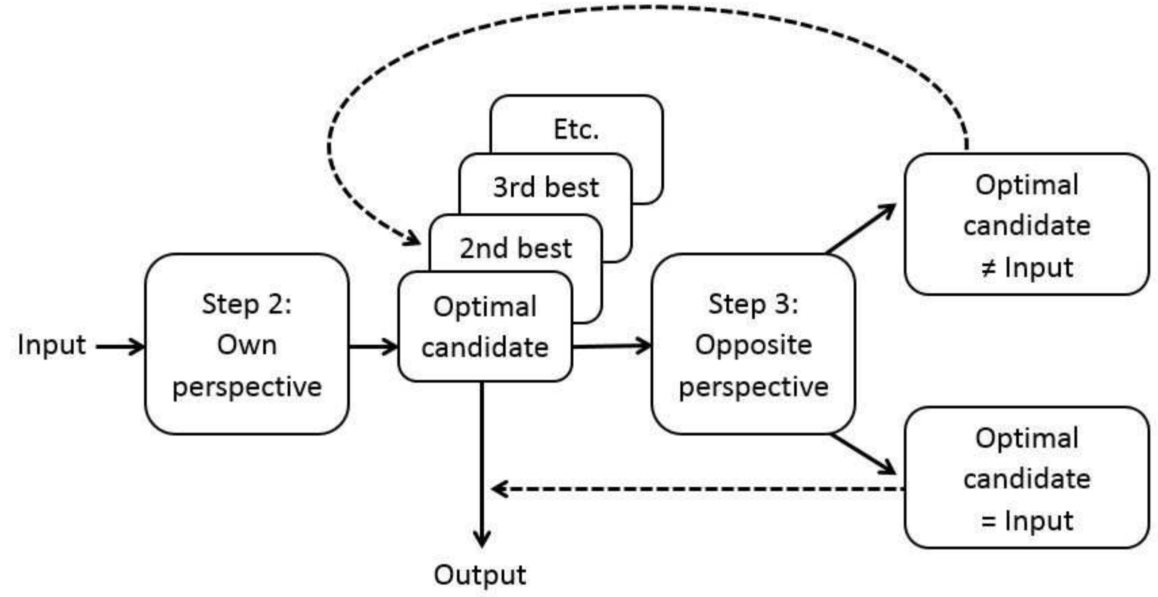


FIGURE 4 | Perspective taking in the ACT-R model of reference. Perspective taking involves selection of the optimal candidate from the language user's own perspective (Step 2), followed by selection of the optimal candidate from the opposite communicative perspective (Step 3). The output is the best candidate in Step 2 that produces the input again in Step 3.

2010). Initially, the model is unable to complete both steps, because this takes too much time. As a consequence, the model is only able to complete the step of the initial selection of a form or meaning and does not take into account the opposite communicative perspective. In ACT-R, processes become more efficient with linguistic experience due to the ACT-R learning mechanism of production compilation (Taatgen and Anderson, 2002). By frequently performing the processes of reference production and comprehension, the relevant production rules are repeatedly carried out in sequence. Production compilation reduces the number of production rules required for processing and hence reduces the amount of time needed for processing. As the model thus gradually gains more processing speed, the model will become more likely to take into account the opposite perspective. Eventually, through the mechanism of production compilation the two-step process of perspective taking may turn into a one-step selection process. Thus, it is predicted that the ability to use perspective taking in real-time conversation is dependent on sufficient processing speed, which in turn is dependent on linguistic experience. Linguistic experience increases with age as well as with frequency of the referring expression: the older the child and the more frequent the referring expression in the language input to the child, the more experience the child can be expected to have with the referring expression.

As perspective taking requires an awareness that speakers may possess different knowledge and make different choices than listeners, the ability of perspective taking in language may be related to the development of a Theory of Mind (ToM). ToM refers to the cognitive capacity to attribute mental states, such as beliefs, desires and intentions, to oneself and others and to understand that the mental states of others may differ from one's own mental states (Premack and Woodruff, 1978). First-order ToM, the capacity to understand what another person thinks, typically emerges in children around the age of 3 or 4 in explicit false-belief tasks (e.g., Wimmer and Perner, 1983). Second-order ToM, which builds on first-order ToM and is the capacity to understand what another person thinks about what yet another person thinks, emerges several years later, around the age of 6 (Perner and Wimmer, 1985). Because of its assumed relation to ToM development, it is conceivable that perspective taking in language only fully develops after age 3 or 4.

The final step of perspective taking is crucial for the mature choice between a pronoun and a full noun phrase (see **Figure 5**). If the model is only able to complete the process of initial selection of a form or meaning and is unable to complete the next process of perspective taking, the model will produce pronouns all the time for expressing anaphoric reference. It will do so even for referents that are not the discourse topic. On the other hand, if the model is able to take into account the opposite perspective of the listener, pronouns are blocked for referents that are not the discourse topic. As a result, the model will restrict its use of pronouns to referents that are the discourse topic. For other discourse referents, the model will select a full noun phrase.

Perspective taking is also crucial for the mature comprehension of pronouns in syntactic binding environments (see **Figure 6**). If the model is unable to complete the process of perspective taking, pronouns in object position in languages such as English will remain ambiguous and can be interpreted as referring to the local subject. However, if the model is able to take into account the opposite perspective of the speaker, this

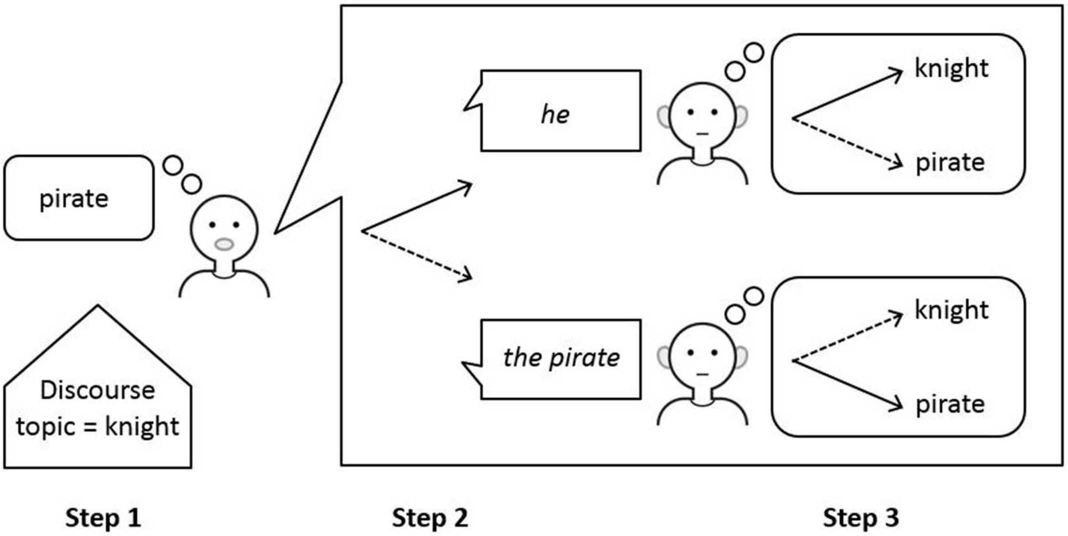


FIGURE 5 | Production of referring expressions in the ACT-R mode of reference. A speaker wishing to refer to a referent that in Step 1 was found not to be the current discourse topic (e.g., the pirate in the final picture of the story in **Figure 1**) preferably uses the pronoun *he* to refer to this referent (indicated by the solid arrow in Step 2). Taking into account the perspective of the listener in Step 3 will reveal that *he* is best interpreted as the current discourse topic (indicated by the solid arrow in the top picture in Step 3). As this referent (knight) is different from the intended referent (pirate), the pronoun is blocked as a potential form and a full noun phrase must be selected. Adapted from Hendriks (2014).

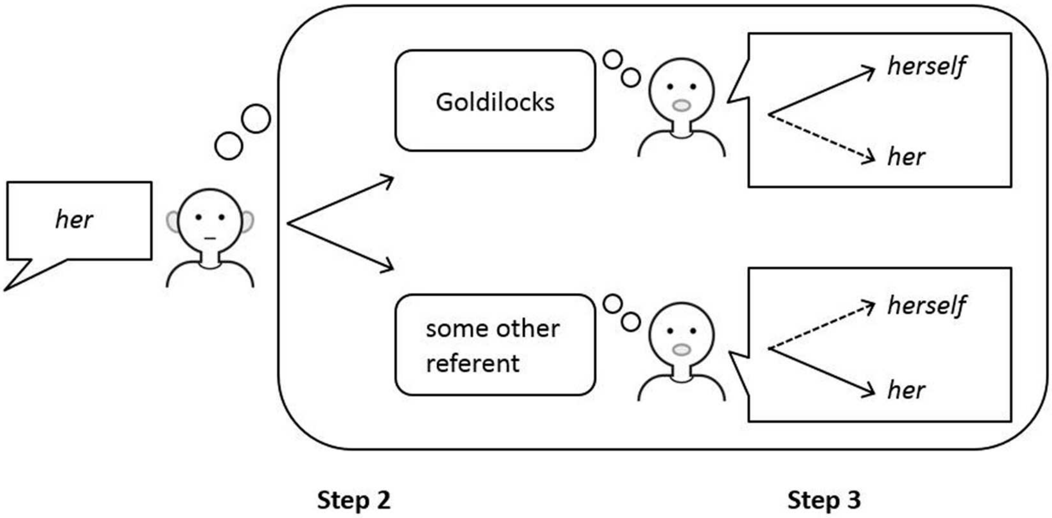


FIGURE 6 | Comprehension of referring expressions in the ACT-R model of reference. A listener hearing the sentence "*Goldilocks washed her*" in a context that provides no referential bias (hence, Step 1 is omitted here) may select Goldilocks or some other referent as the antecedent of the pronoun *her* (indicated by the two solid arrows in Step 2) and will make a random choice. If Goldilocks is selected as the antecedent of *her*, taking into account the perspective of the speaker in Step 3 will reveal that reference to Goldilocks is best expressed by the reflexive *herself* (indicated by the solid arrow in the top picture in Step 3). As this form is different from the heard form *her*, the referent Goldilocks is blocked as a potential antecedent for the pronoun *her* and some other referent must be selected. Adapted from Hendriks (2014).

bound interpretation will be blocked and pronouns in object position will be interpreted as being non-coreferential with the local subject.

Thus, perspective taking in production and comprehension has the effect of avoiding misunderstanding between speaker and listener. Note that avoiding misunderstanding crucially differs from avoiding ambiguity: producing referentially ambiguous expressions such as pronouns is permitted by the model, as long as the ambiguity does not result in misunderstanding between speaker and listener in the given discourse context. Our view of perspective taking as a crucial step in the production and comprehension of particular linguistic expressions that is nevertheless still difficult for children, may contribute to the current debate about the role of perspective taking in language. Various positions have been put forward in this debate: that language users are initially egocentric and only adjust their perspective when circumstances demand (e.g., Horton and Keysar, 1996; Keysar et al., 2003), that language users are initially egocentric because they are unable to fully discount their own perspective (Barr, 2008), and that perspective taking is one of many cues in language processing that is used early on (e.g., Brown-Schmidt et al., 2008). As pointed out by Brown-Schmidt (2009), empirical findings have been equivocal about the online use of perspective taking in language processing, so one of the main challenges for models of perspective taking in language is to account for why perspective taking sometimes constrains online processing and sometimes does not.

Because of its direct appeal to cognitive capacities such as WM, processing speed and ToM, which can vary among individual speakers and listeners, the ACT-R model of reference seems particularly suited to explain and predict different patterns of variation in reference production and reference comprehension. In the next section, the model's predictions are discussed for speakers' underspecification and overspecification of referring expressions and for listeners' incorrect or overly liberal interpretation of referring expressions.

EXPLAINING AND PREDICTING INDIVIDUAL VARIATION

Underspecification of Referring Expressions

As mentioned above, children prefer to use pronouns over full noun phrases, even when referring to referents that are not the discourse topic (e.g., Karmiloff-Smith, 1985). As these pronouns underspecify their referent, they may cause misunderstanding for a listener. For this reason, adult speakers generally use full noun phrases when referring to a referent that is not the discourse topic.

Performing simulations with the ACT-R model of reference, we can investigate the effects of cognitive factors on reference production and comprehension by manipulating features of the model. van Rij (2012, Chap. 3) modeled the production of referring expressions in a linguistic discourse and investigated the effects of WM on the performance of the model. The performance of two variants of this model was compared: a model with a low WM capacity and a model with a high WM capacity (implemented by spreading activation). In the simulations run with the models, the models were presented with stories of five sentences each about two referents of the same gender, similar to the narratives produced by children and adults for the picture story in **Figure 1**. Each story started with the first referent being the subject of the sentence and hence the topic of the discourse. Halfway through the story, the topic shifted from the first to the second referent by making this second referent the subject of that sentence and the next one. After having presented the model with this linguistic discourse, the task of the models was to produce a referring expression to refer back to the first referent, which at that point in the discourse was not the topic anymore.

The low WM capacity model produced underspecified pronouns to refer back to the first referent in 86% of the cases, and produced explicit full noun phrases in the remaining 14% of the cases. In contrast, the high WM capacity model produced pronouns in only 11% of the cases and full noun phrases in the large majority of cases, namely 88%. The performance of the low WM capacity model reflects the performance of the children in the study by Hendriks et al. (2014). They tested 4- to-7-year-old Dutch-speaking children on a narrative elicitation task based on picture stories consisting of six pictures each, such as in **Figure 1**. The stories elicited a topic shift from the first referent (the pirate) to the second referent (the knight) halfway through the story. To re-introduce the first referent in the final picture of the story, which is not the discourse topic anymore at the moment of re-introduction, the children in the study produced pronouns in 62% of the cases and full noun phrases in 38% of the cases. This preference for pronouns was in accordance with the performance of the low WM capacity model. In contrast, the performance of the high WM capacity model reflects the performance of the young Dutch adults in the same study. These adults produced pronouns in only 9% of the cases and produced full noun phrases in 91% of the cases.

The model's prediction that WM is a crucial factor in the choice between a pronoun and a full noun phrase is supported by experimental evidence from narrative elicitation studies with various populations. Hendriks et al. (2014) found a positive correlation between the use of full noun phrases by the children and their scores on an auditory memory task (word repetition): the higher the children's memory scores, the more often they used a full noun phrase to re-introduce the first referent. A similar effect of WM (this time measured by an n-back task) was found in a study with children with autism, children with ADHD and typically developing children in the age range between 6 and 12 years old who were tested on the same narrative elicitation task (Kuijper et al., 2015). In addition to the effect of WM, Kuijper et al. also found an effect of second-order ToM: the higher the children's scores on second-order ToM (as measured by a false-belief task), the more full noun phrases they used for referring back to the first referent.

Inspecting the ACT-R model's performance allows us to more closely examine the reasons for selecting an underspecified pronoun, in particular for the way WM and processing speed influence this choice. Of the 86% of cases in which the low WM capacity model produced a pronoun, in two third of these cases (57%) this is caused by a low amount of spreading activation (van Rij, 2012, p. 64). Due to its low amount of spreading activation, the low WM capacity model does not take into account the grammatical roles of referents in the local discourse. It only relies on frequency and recency of mentioning. Hence, it shows a strongly reduced preference for selecting the subject of the previous sentence (the second referent) as the discourse topic. As a result, the low WM capacity model is not very accurate in determining the discourse topic and will often select the first referent. If the model incorrectly selects the first referent as the discourse topic, a pronoun is the optimal form for re-introducing this first referent, both according to the linguistic constraints and after considering the opposite perspective. However, for a listener who has access to the preceding linguistic discourse and correctly selects the second referent as the discourse topic, this pronoun will be interpreted as referring to the second referent. Thus, the use of a pronoun after a topic shift will result in misunderstanding.

In addition to low WM capacity, the ACT-R model reveals a second reason for using an underspecified pronoun after a topic shift, namely insufficient speed of sentence processing. Of the 86% of cases in which the low WM capacity model produced a pronoun, in one third of these cases (29%) this is caused by insufficient processing speed (van Rij, 2012, p. 64). The linguistic constraints lead the model to have a general preference for using a pronoun. Only if the model succeeds in taking into account the listener's perspective to check the recoverability of the initially selected form will the model block the use of a pronoun and select a more explicit full noun phrase instead (see **Figure 5**). As the completion of the process of perspective taking is dependent on sufficient processing speed (van Rij et al., 2010), insufficient processing speed results in the use of underspecified pronouns. As we saw above, also less advanced ToM abilities are related to the use of underspecified pronouns (Kuijper et al., 2015). It is conceivable that children with less advanced ToM abilities are slower in perspective taking, thus having insufficient time to complete the process of perspective taking during online sentence processing.

So, based on simulations of the ACT-R model, it can be argued that the mature use of pronouns requires sufficient WM capacity (which increases through maturation) and sufficient processing speed (which increases through linguistic experience). The observed differences between children and adults in their production of referring expressions can thus be explained by individual differences in their WM capacity and processing speed.

In addition to explaining existing data regarding reference production, the ACT-R model also generates novel predictions, that can be tested in subsequent experiments. For example, adults with a low WM capacity but sufficient processing speed are predicted to frequently use a pronoun to refer to a referent which they incorrectly take to be the discourse topic, just like children. This follows from their expected failure to use grammatical role information from the previous sentence in determining the discourse topic. Indeed, the elderly adults with a mean age of almost 80 in the study of Hendriks et al. (2014), whose average score on the memory task was significantly lower than that of the young adults, but who can be expected to still have sufficient processing speed, produced pronouns to re-introduce the first referent in almost half of the cases (47%). An indication that the elderly adults' production of underspecified forms is due to their low WM capacity, and is not caused by their failure in perspective taking, is the observation that they produced significantly more full noun phrases when re-introducing the first referent in the sixth picture than when referring to the second referent in the fifth picture. This suggests that elderly adults do take the listener into account when re-introducing the first referent (Hendriks et al., 2014), and thus possess sufficient processing speed.

Overspecification of Referring Expressions

Another novel prediction of the ACT-R model of reference, one that has not been tested yet, is that particular linguistic discourse contexts lead speakers to produce overly specific referring expressions because of insufficient WM capacity. That is, in these discourse contexts speakers with insufficient WM capacity but sufficient processing speed are expected to use a full noun phrase to refer to the discourse topic, although a pronoun would have sufficed.

Insufficient WM capacity can occur for several reasons. Speakers may have insufficient WM capacity due to age or cognitive deficits. Alternatively, they may have insufficient WM capacity available because their WM is overloaded by other cognitive processes. This can happen when a speaker has to carry out two or more tasks simultaneously. In ACT-R, the effect of high cognitive load is similar to the effect of low WM capacity (van Rij et al., 2013): due to the mechanism of spreading activation, goal-relevant information spreads activation to other chunks in declarative memory. If the number of sources from which activation is spread increases, the amount of spreading activation received by individual chunks decreases, because the total amount of spreading activation is fixed. In a situation of high cognitive load, more information needs to be maintained in an activated state. Thus, more sources spread the fixed amount of spreading activation. As a result, the subject of the previous sentence spreads less activation to the discourse referent associated with the subject and hence this referent is less likely to be selected as the discourse topic.

The particular linguistic discourse contexts that are predicted to lead speakers to overspecify their referring expressions are contexts in which the discourse topic shifts from a more frequently or more recently mentioned referent to a less frequently or less recently mentioned referent. In such discourse contexts, a speaker with sufficient WM capacity will signal the topic shift (e.g., by expressing the new topic as a full noun phrase in subject position) and then continue to refer to this new topic using a pronoun. A speaker with insufficient WM capacity, on the other hand, may continue to refer to this new topic by using a full noun phrase. This is because the lack of sufficient WM capacity causes the speaker to rely less on grammatical role information from the previous sentence and more on frequency and recency of mentioning when determining the discourse topic. If the new topic is a less frequently or less recently mentioned referent, the speaker may incorrectly assume that this referent has not been established as the discourse topic yet and use a full noun phrase to refer to this referent.

Thus, it is predicted that young adult speakers carrying out an additional task that taxes their WM will produce more overspecified referring expressions for reference to the new topic immediately after a topic shift. The speaker will continue to use such overspecified referring expressions until the accessibility of the new discourse topic has increased by frequent mentioning or recency. Some suggestive evidence in favor of this prediction comes from the use of referring expressions by adult learners of a second language. Speaking a foreign language generally requires more cognitive resources, including WM, than speaking the native language (e.g., Linck et al., 2013). In a study investigating narratives produced by adult intermediate and advanced learners of French and English when retelling a silent cartoon movie, the adult second-language learners were found to overspecify referring expressions compared to native speakers of the two languages and to use definite noun phrases where pronouns could be used (Leclercq and Lenart, 2013). These overspecifications particularly occurred when the speakers had to re-introduce the second character. As the two characters appeared to be of a different gender to many of the participants, a pronoun would have sufficed. Although Leclercq and Lenart explain the overspecification of referring expressions in second-language learners as a conscious risk-avoiding strategy, these overspecifications may very well be the unconscious effects of insufficient WM capacity.

In addition to variation in production, in particular with respect to overspecification and underspecification, the ACT-R model also explains variation in comprehension. Below, we discuss some of the variation observed in adults' and children's interpretations of pronouns.

Incorrect Interpretation of Referring Expressions

Adults who have less WM capacity available because their WM is taxed by an additional task are also more likely to select an incorrect referent for a pronoun. In particular, they are predicted to show difficulty comprehending a topic shift. As they are less likely to use grammatical role information from the preceding sentence due to the low amount of spreading activation, they will solely rely on frequency and recency of mentioning of the referents. In case of a topic shift, this will often result in selection of the incorrect referent as the discourse topic and hence as the antecedent of the pronoun.

In a dual-task experiment, van Rij et al. (2013) tested this prediction of the ACT-R model. Adult participants had to perform two tasks at the same time. The linguistic task was a pronoun comprehension task. The additional task was the memorization of a series of digits. While memorizing either three digits (low cognitive load condition) or six digits (high cognitive load condition), participants read short stories consisting of four sentences. The stories featured two referents of the same gender, which were only referred to with proper names. The final sentence started with a potentially ambiguous subject pronoun that could in principle refer to both referents (e.g., "*He has played soccer for twenty years*"). Following the story, participants received a comprehension question, which asked for the referent of the pronoun. They received two types of stories: stories with and stories without a topic shift. These two types of stories only differed in the grammatical roles of the referents in the second and third sentence (subject or non-subject). In the topic-shift stories, the topic was shifted from the most frequently mentioned referent to the other referent by making this other referent the subject of the second and third sentence. In the non-topic-shift stories, the most frequently mentioned referent remained the subject of the sentence throughout the story. After answering the comprehension question, participants had to type in the digits that were presented to them before the start of the story. Each participant was tested in both cognitive load conditions.

As predicted, adults less often selected the subject of the previous sentence as the referent of the pronoun in the high cognitive load condition than in the low cognitive load condition. Instead, they selected the most frequently mentioned other referent. This effect of cognitive load was limited to stories with a topic shift and did not affect stories without a topic shift, which was in line with the predictions of the model. Thus, adult listeners more often assign an incorrect interpretation to an anaphoric pronoun under high cognitive load and select the most frequent referent in the discourse, instead of the linguistically most prominent referent.

Overly Liberal Interpretation of Referring Expressions

The predictions discussed above mainly concerned WM capacity. An exception was the prediction that, in addition to low WM capacity, also insufficient speed of sentence processing results in the production of underspecified pronouns after a topic shift. Here, we discuss another prediction of the ACT-R model concerning processing speed, namely the prediction that insufficient processing speed results in an overly liberal interpretation of object pronouns in syntactic binding environments. Note that in the case of object pronouns, low WM capacity is predicted not to have an effect, as the correct interpretation of object pronouns is independent of the linguistic discourse.

As mentioned above, to interpret the object pronoun *her* in the sentence "*Goldilocks washed her*" and to restrict its interpretation to a referent that is not the local subject, a listener must take into account the perspective of the speaker. Taking into account the opposite perspective in addition to one's own perspective is expected to take more time than only considering one's own perspective, as the process of perspective taking is modeled in the ACT-R model of reference as two consecutive processes of optimization (Hendriks et al., 2007). Therefore, sufficient processing speed is needed to complete the process of perspective taking. Processing speed is increased by the ACT-R learning mechanism of production compilation, which depends on linguistic experience. Children may not have sufficient processing speed yet to be able to complete the process of perspective taking within the alotted time. However, they may be able to take into account the opposite perspective when given more time for interpretation.

In the ACT-R model, new words arrive at a fixed rate. This rate cannot be influenced by the listener. Therefore, time for interpretation of a word is limited to the time until the arrival of the next word. If the time until the next word is increased, children have more time for the interpretation of a sentence-internal pronoun and may be able to complete the process of perspective taking more often. van Rij et al. (2010) carried out a picture verification task with 4- to 7-year-old Dutch children to test this prediction. The children received sentences such as "*The bear is tickling him with a feather*" and had to say whether the sentence matched an accompanying picture or not. As the pronoun occurs mid-sentence, time for interpretation is limited to the arrival of the next word. Half of the sentences were presented to the child at a normal speech rate and the other half at a slower speech rate of 2/3 of normal speech rate. In the slower speech rate condition, the children had more time for the interpretation of the pronoun due to the extra time between words.

van Rij et al. (2010) found that, if children displayed the Delay of Principle B Effect, slowing down the speech rate improved their comprehension of pronouns. In contrast, slow speech rate had a negative effect on their (already adult-like) comprehension of reflexives. These selective beneficial effects of slowed-down speech support the assumption implemented in the ACT-R model that the mature interpretation of object pronouns requires perspective taking. It is also consistent with the view of perspective taking as an online and local process, rather than a pragmatic and end-of-sentence process, as it is dependent on sufficient processing speed during online sentence processing.

Based on these outcomes, it is further predicted that the mature comprehension of object pronouns is related to advanced ToM abilities, in the same way that avoiding to produce underspecified pronouns after a topic shift is related to advanced ToM abilities, as both processes are hypothesized to require perspective taking (see **Figures 5** and **6**). This contrasts with the mature comprehension of reflexives and the mature production of pronouns in topic continuation situations, which are expected not to be dependent on the additional step of perspective taking, as the linguistic constraints already lead to the correct output in these cases (Hendriks, 2014).

Another prediction that can be experimentally tested is that features of the linguistic discourse influence the offline interpretation of object pronouns of listeners with low processing speed, but not of listeners with high processing speed. Perspective taking generates internal feedback, which restricts the interpretation of object pronouns and makes this process less dependent on the discourse (see **Figure 6**). As perspective taking depends on processing speed, reduction of the influence of the discourse also depends on processing speed. Thus, it is expected that the linguistic discourse influences children's offline interpretation of object pronouns, but not adults' (although it may influence adults' online processing). In particular, if the *correct* antecedent of the pronoun is the most frequently and most recently mentioned referent, children will be biased toward the correct antecedent in Step 2, whereas if an *incorrect* antecedent is the most frequently and most recently mentioned referent, children will be biased toward the incorrect antecedent in Step 2. Without the internal feedback provided by perspective taking in Step 3, children will stick to their initial choice made in Step 2. Children's overreliance on the linguistic discourse is predicted to disappear with increasing processing speed.

DISCUSSION

In this paper, it was shown how computational modeling of referential choice within the cognitive architecture ACT-R can yield more insight into the cognitive mechanisms underlying the observed variation in speakers' production and listeners' comprehension of referring expressions. The ACT-R model of reference uses general memory principles of ACT-R to build a representation of the linguistic discourse, employs linguistic constraints to make an initial selection of a form or a meaning in that discourse, and performs perspective taking to check whether this initially selected form or meaning will indeed allow for mutual understanding between speaker and listener in the given discourse context. The mechanism of perspective taking does not require any additional assumptions in ACT-R. Rather, it comes for free to the model because comprehension is implemented according to the same principles of optimization as production. Perspective taking is merely modeled as the addition of an extra step of optimization from the opposite communicative perspective.

Because the ACT-R model of reference is based on verified assumptions about human cognition, the model is able to explain some of the observed variation in referential choice from variation in speakers' and listeners' cognitive capacities. The cognitive processes required for a specific referential task may exceed the cognitive capacities of some speakers and listeners, but not of others. Also, the cognitive processes required for some referential tasks may be more demanding than those for other tasks. Individual variation in the cognitive capacities of speakers and listeners and limitations in these capacities thus give rise to variation among and within individuals and across tasks. A first process that is expected to require sufficient cognitive capacities is the construction and maintenance of a representation of the linguistic discourse. This process is predicted by the ACT-R model to depend on the availability of sufficient WM capacity. Another process that is expected to be effortful, as it requires an additional step in production and comprehension, is perspective taking. The ACT-R model predicts that perspective taking depends on sufficient processing speed.

Hence, one source of variation in referential choice is WM capacity. Low WM capacity is argued to lead to difficulty in taking discourse prominence into account in building a representation of the linguistic discourse. Therefore, low WM capacity is expected to be involved in the production of underspecified referential forms (cf. Vogels et al., 2014). This explains why children and elderly adults occasionally produce pronouns without a clear reference in their narratives. Furthermore, an incorrect representation of the linguistic discourse due to low WM capacity is predicted to result in errors in the interpretation of pronouns as well. This explains children's difficulty in determining the correct referent of a pronoun after a topic shift as well as adults' child-like pattern of pronoun interpretation when their WM is taxed by an additional task.

Another source of variation in referential choice is processing speed. Insufficient speed of sentence processing is predicted to lead to a failure to consider the opposite communicative perspective. This is argued to explain children's production of unrecoverable pronouns in narratives as well as their overly liberal interpretation of pronouns in object position. This explanation of children's non-adult-like referential choices is in line with the view that perspective taking initially is an effortful process that requires the adjustment of one's own perspective (e.g., Epley et al., 2004; Barr, 2008). In adults, due to the ACT-R learning mechanism of production compilation, the two-step process of perspective taking may be reduced to a one-step selection process. This could result in perspective taking processes becoming automatic and occurring early in adults (cf. Brown-Schmidt et al., 2008; Brennan and Hanna, 2009).

In recent years, several probabilistic approaches have been proposed in order to account for variation in referential choice (e.g., Frank and Goodman, 2012; van Gompel et al., 2012; Mitchell et al., 2013). For example, Frank and Goodman (2012) assume that listeners interpret referring expressions as a function of the prior probability that an object would be referred to and the probability that the speaker would use a particular word to refer to this object. In their approach, speakers are rational agents who choose words that are informative in context and reduce uncertainty about the referent. This view is criticized by Gatt et al. (2013), who argue that the observation of overspecification by human speakers provides evidence that speakers may not be rational agents after all. Like Frank and Goodman's model, the ACT-R model of reference presented here also includes probabilistic processes, and furthermore assumes that speakers and listeners are rational agents. However, in contrast to Frank and Goodman's model, in the ACT-R model of reference perfect rationality is not always achieved by speakers and listeners due to limitations in their cognitive capacities. Because of its bounded rationality, the ACT-R model occasionally gives rise to overspecification and underspecification.

The ACT-R model of reference is not specifically geared toward one task, but is based on general principles of human information storage, retrieval and processing in combination with general linguistic constraints on reference. Hence, the model not only explains existing data, but is also able to generate novel predictions. For example, the model predicts that speakers who are under cognitive load will produce more overly specific referring expressions after a topic shift. Also, it predicts that listeners without sufficient processing speed will be influenced in their interpretation of object pronouns by the frequencies of referents in the linguistic discourse. These predictions can be tested in new psycholinguistic experiments and in other referential tasks, providing further evidence on how referential choice varies across different tasks and among and within individuals. For example, very little is known yet about the decline of referential abilities in healthy elderly adults and how this relates to their cognitive abilities. Is it true that elderly adults' changing performance on many cognitive tasks (including referential tasks) does not reflect cognitive decline, but instead reflects increased knowledge and corresponding memory search demands, as Ramscar et al. (2014) argue? Cognitive modeling could help to answer this question. Also, in addition to children and adults with autism or ADHD, other clinical populations could be studied that have been suggested to have limitations in their WM capacity, processing speed or both, such as patients with Alzheimer's disease, Broca's aphasia, or multiple sclerosis (e.g., Almor et al., 1999; Love et al., 2001; Piñango and Burkhardt, 2001). Studying these populations through cognitive modeling could reveal more about reference processing in general as well as about the cognitive deficits in these clinical populations.

The task of the ACT-R model is to make a choice between pronouns and definite descriptions in production, and between different discourse referents for a pronoun in comprehension. Most computational models for the generation of referring expressions, in contrast, focus on the choice between different definite descriptions for a particular referent, such as between *the grey desk*, *the desk facing left* and *the gray desk facing left* (Mitchell et al., 2013). To obtain a more comprehensive model of referential choice, the ACT-R model discussed here should be extended to allow for these specific referential choices between different definite descriptions as well. One proposal is by Guhe (2012), who modeled human behavior in the so-called iMAP task in ACT-R. In this task, participants had to reproduce a route on a map by referring to landmarks on the map using features such as color, number and kind (e.g., *red bugs*, *four bugs*, or *four red bugs*). Guhe developed two ACT-R models of human behavior in this task, the first one an extension of the incremental algorithm of Dale and Reiter (1995) and the second one based on a fixed template of features. Both cognitive models select features on the basis of the utility of the corresponding ACT-R production rule: production rules contributing to a successful interaction are selected with a higher probability. Furthermore, both cognitive models are able to adapt the utility value of features to feedback of whether a referring expression was used successfully. While the second model had a higher correlation with the human data, it was more geared toward the specific task. On the other hand, the first model was more general, but had difficulty predicting under- and overspecified referring expressions because of its goal to generate a uniquely distinguishing expression (for discussion, see Guhe, 2012, p. 320). However, it may be possible to circumvent this problem of the first model by replacing the goal of generating a uniquely distinguishing expression by the goal of finding a bidirectionally optimal expression, as in the ACT-R model presented here, thus aiming at avoiding misunderstanding rather than avoiding ambiguity. As the two models capture the general patterns of adaptive change in referential choice that are observed in the human data, they illustrate that modeling specific referential choices between different definite descriptions is in principle possible in ACT-R.

Another useful addition to the model would be the inclusion of visual factors in the calculation of referent activation, as the presence or absence of characters in the visual context also affects the choice between a pronoun and a full noun phrase (Fukumura et al., 2010). This would also be more in line with the view that referents are bundles of multimodal features (e.g., van der Sluis and Krahmer, 2007; van der Sluis et al., 2008), which led van der Sluis and Krahmer (2007) to include various types of pointing gestures in their computational model for the generation of referring expressions. A more realistic calculation of referent activation would also require the inclusion of further linguistic and cognitive factors, such as coherence relations between utterances, animacy, and first mention. As the first two factors have been incorporated as constraints in linguistic analyses (e.g., de Hoop, 2013; Hendriks, 2014), they may alternatively be included in the set of linguistic constraints implemented in the ACT-R model.

Despite these limitations, the ACT-R model of referential choice seems to be a promising starting point for the further exploration of factors involved in referential choice. By running computational simulations that manipulate the cognitive and linguistic factors implemented in the ACT-R model, the model can generate quantitative predictions about performance in various populations of speakers and listeners on a variety of referential tasks. These predictions can be tested experimentally, thus allowing us to gain further insights in the dependence of referential choice on these cognitive and linguistic factors.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This investigation was financially supported by a Vici grant from the Netherlands Organisation for Scientific Research awarded to PH (grant no. 277-70-005).

ACKNOWLEDGMENT

The author thanks Jacolien van Rij for discussions on the ACT-R model and for her valuable comments on an earlier version of this paper.

REFERENCES

Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., and Tyler, L. K. (1999). Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's Disease. *Brain Lang.* 67, 202–227. doi: 10.1006/brln.1999.2055

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York, NY: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Ariel, M. (1988). Referring and accessibility. *J. Linguist.* 24, 65–87. doi: 10.1017/S0022226700011567

Ariel, M. (1990). *Anaphoric Antecedents*. London: Croom Helm.

Arnold, J. (2010). How speakers refer: the role of accessibility. *Lang. Linguist. Compass* 4, 187–203. doi: 10.1111/j.1749-818x.2010.00193.x

Arnold, J., Bennetto, L., and Diehl, J. J. (2009). Reference production in young speakers with and without autism: effects of discourse status and processing constraints. *Cognition* 110, 131–146. doi: 10.1016/j.cognition.2008.10.016

Barr, D. J. (2008). Pragmatic expectations at linguistic evidence: listeners anticipate but do not integrate common ground. *Cognition* 109, 18–40. doi: 10.1016/j.cognition.2008.07.005

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *J. Semant.* 17, 189–216. doi: 10.1093/jos/17.3.189

Brennan, S. E., and Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics Cogn. Sci.* 1, 274–291. doi: 10.1111/j.1756-8765.2009.01019.x

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychon. Bull. Rev.* 16, 893–900. doi: 10.3758/PBR.16.5.893

Brown-Schmidt, S., Gunlogson, C., and Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition* 107, 1122–1134. doi: 10.1016/j.cognition.2007.11.005

Budiu, R., and Anderson, J. R. (2004). Interpretation-based processing: a unified theory of semantic sentence comprehension. *Cogn. Sci.* 28, 1–44. doi: 10.1207/s15516709cog2801_1

Chien, Y. C., and Wexler, K. (1990). Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Lang. Acquisit.* 1, 225–295. doi: 10.1207/s15327817la0103_2

Chomsky, N. (1981). *Lectures on Government and Binding*: *The Pisa Lectures*. Berlin: Mouton de Gruyter.

Daily, L. Z., Lovett, M. C., and Reder, L. M. (2001). Modeling individual differences in working memory performance: a source activation account. *Cogn. Sci.* 25, 315–353. doi: 10.1207/s15516709cog2503_1

Dale, R., and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902_3

Dale, R., and Viethen, J. (2010). "Attribute-centric referring expression generation," in *Empirical Methods in Natural Language Generation*, eds E. Krahmer and M. Theune (Berlin: Springer Verlag), 163–179.

de Hoop, H. (2013). Incremental optimization of pronoun interpretation. *Theor. Linguist.* 39, 87–93. doi: 10.1515/tl-2013-0005

De Villiers, J., Cahillane, J. and Altreuter, E. (2006). "What can production reveal about Principle B?," in *Proceedings of the Inaugural Conference on Generative Approaches to Language Acquisition–North America*, eds K. U. Deen, J. Nomura, B. Schulz, and B. D. Schwartz (Mansfield, CT: University of Connecticut, Occasional Papers in Linguistics 4), 89–100.

Deutsch, W., and Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition* 11, 159–184. doi: 10.1016/0010-0277(82) 90024-5

Engelhardt, P. E., Bailey, K. G. D., and Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *J. Mem. Lang.* 54, 554–573. doi: 10.1016/j.jml.2005.12.009

Epley, N., Morewedge, C. K., and Keysar, B. (2004). Perspective taking in children and adults: equivalent egocentrism but differential correction. *J. Exp. Soc. Psychol.* 40, 760–768. doi: 10.1016/j.jesp.2004.02.002

Frank, M., and Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science* 336:998. doi: 10.1126/science.1218633

Fukumura, K., van Gompel, R., and Pickering, M. J. (2010). The use of visual context during the production of referring expressions. *Q. J. Exp. Psychol.* 63, 1700–1715. doi: 10.1080/17470210903490969

Fukumura, K., and van Gompel, R. P. G. (2010). Choosing anaphoric expressions: do people take into account likelihood of reference? *J. Mem. Lang.* 62, 52–66. doi: 10.1016/j.jml.2009.09.001

Gatt, A., van Gompel, R. P. G., van Deemter, K., and Krahmer, E. (2013). "Are we Bayesian referring expression generators?," in *Proceedings of the PRE-CogSci 2013 Workshop on the Production of Referring Expressions: Bridging the Gap Between Cognitive and Computational Approaches to Reference*, Berlin.

Givón, T. (ed.) (1983). *Topic Continuity in Discourse. A Quantitative Cross-Language Study*. Amsterdam: John Benjamins.

Grimshaw, J., and Samek-Lodovici, V. (1998). "Optimal subjects and subject universals," in *Is the Best Good Enough?*, eds P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis, and D. Pesetsky (Cambridge, MA: MIT Press), 193–219.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.* 21, 203–225.

Guhe, M. (2012). Utility-based generation of referring expressions. *Topics Cogn. Sci.* 4, 306–329. doi: 10.1111/j.1756-8765.2012.01185.x

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307. doi: 10.2307/416535

Gundel, J. K., Hedberg, N., and Zacharski, R. (2012). Underspecification of cognitive status in reference production: aome empirical predictions. *Topics Cogn. Sci.* 4, 249–268. doi: 10.1111/j.1756-8765.2012.01184.x

Hendriks, P. (2014). *Asymmetries Between Language Production and Comprehension. Studies in Theoretical Psycholinguistics*, Vol. 42. Dordrecht: Springer.

Hendriks, P., Englert, C., Wubs, E., and Hoeks, J. C. J. (2008). Age differences in adults' use of referring expressions. *J. Logic Lang. Inform.* 17, 443–466. doi: 10.1080/13825585.2011.607228

Hendriks, P., Koster, C., and Hoeks, J. C. J. (2014). Referential choice across the lifespan: why children and elderly adults produce ambiguous pronouns. *Lang. Cogn. Neurosci.* 29, 391–407. doi: 10.1080/01690965.2013.766356

Hendriks, P., van Rijn, H., and Valkenier, B. (2007). Learning to reason about speakers' alternatives in sentence comprehension: a computational account. *Lingua* 117, 1879–1896. doi: 10.1016/j.lingua.2006.11.008

Horton, W. S., and Keysar, B. (1996). When do speakers take into account common ground? *Cognition* 59, 91–117. doi: 10.1016/0010-0277(96)81418-1

Karmiloff-Smith, A. (1985). Language and cognitive processes from a developmental perspective. *Lang. Cogn. Process.* 1, 61–85. doi: 10.1080/01690968508402071

Keysar, B., Lin, S., and Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition* 89, 25?–41. doi: 10.1016/S0010-0277(03)00064-7

Kibrik, A. A., Khudyakova, M. V., Dobrov, G. B., and Linnik, A. S. (2013). "Referential choice: a cognitively based modeling study," in *Proceedings of the PRE-CogSci 2013 Workshop on the Production of Referring Expressions: Bridging the Gap between Cognitive and Computational Approaches to Reference*, Berlin.

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Krahmer, E., and van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI_a_00088

Kuijper, S. J. M., Hartman, C. A., and Hendriks, P. (2015). Who is he? Children with ASD and ADHD take the listener into account in their production of ambiguous pronouns. *PLoS ONE* 10:e0132408. doi: 10.1371/journal.pone. 0132408

Leclercq, P., and Lenart, E. (2013). Discourse cohesion and accessibility of referents in oral narratives: a comparison of L1 and L2 acquisition of French and English. *Discours* 12. doi: 10.4000/discours.8801

Lewis, R. L., and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419. doi: 10.1207/s15516709cog0000_25

Linck, J. A., Osthus, P., Koeth, J. T., and Bunting, M. F. (2013). Working memory and second language comprehension and production: a meta-analysis. *Psychon. Bull. Rev.* 21, 861–883. doi: 10.3758/s13423-013-0565-2

Love, T., Swinney, D., and Zurif, E. (2001). Aphasia and the time-course of processing long distance dependencies. *Brain Lang.* 79, 169–170.

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., and Reape, M. (2006). A reference architecture for natural language generation systems. *Nat. Lang. Eng.* 12, 1–34. doi: 10.1017/S1351324906004104

Misker, J. M. V., and Anderson, J. R. (2003). "Combining optimality theory and a cognitive architecture," in *Proceedings of the Fifth International Conference on Cognitive Modeling*, eds F. Detje, D. Dörner, and H. Schaub (Bamberg: Universitäts-Verlag Bamberg), 165–170.

Mitchell, M., van Deemter, K., and Reiter, E. (2013). "Generating expressions that refer to visible objects," in *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, (Stroudsburg, PA: The Association for Computational Linguistics).

Perner, J., and Wimmer, H. (1985). John thinks that mary thinks that - attribution of 2nd-order beliefs by 5-year-old to 10-year-old children. *J. Exp. Child Psychol.* 39, 437–471. doi: 10.1016/0022-0965(85)90051-7

Piñango, M. M., and Burkhardt, P. (2001). Pronominals in Broca's aphasia comprehension: the consequences of syntactic delay. *Brain Lang.* 79, 167–168.

Premack, D. G., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1, 515–526. doi: 10.1017/S0140525X00076512

Prince, A., and Smolensky, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell.

Ramscar, M., Hendrix, P., Shaoul, C., Millin, P., and Baayen, H. (2014). The myth of cognitive decline: non-linear dynamics of lifelong learning. *Topics Cogn. Sci.* 6, 5–42. doi: 10.1111/tops.12078

Reitter, D., Keller, F., and Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cogn. Sci.* 35, 587–637. doi: 10.1111/j.1551-6709.2010.01165.x

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguist. Inq.* 27, 720–731.

Song, H., and Fisher, C. (2005). Who's "she?" Discourse prominence influences preschoolers' comprehension of pronouns. *J. Mem. Lang.* 52, 29–57. doi: 10.1016/j.jml.2004.06.012

Spenader, J., Smits, E., and Hendriks, P. (2009). Coherent discourse solves the pronoun interpretation problem. *J. Child Lang.* 36, 23–52. doi: 10.1017/S0305000908008854

Taatgen, N. A., and Anderson, J. R. (2002). Why do children learn to say "broke?" A model of learning the past tense without feedback. *Cognition* 86, 123–155. doi: 10.1016/S0010-0277(02)00176-2

van Deemter, K., Gatt, A., van Gompel, R. P. G., and Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Topics Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01187.x

van der Sluis, I., and Krahmer, E. (2007). Generating multimodal references. *Discour. Process.* 44, 145–174. doi: 10.1016/j.ejrad.2011.08.005

van der Sluis, I., Piwek, P., Gatt, A., and Bangerter, A. (2008). "Multimodal referring expressions in dialogue," in *Proceedings of the Symposium on Multimodal Output Generation (MOG 2008) Held at the AISB 2008 Convention*. Aberdeen.

van Gompel, R., Gatt, A., Krahmer, E., and van Deemter, K. (2012). "PRO: a computational model of referential overspecification," in *Proceedings of the Architectures and Mechanisms for Language Processing (AMLaP) Conference* (Trento: University of Trento).

van Rij, J. (2012). *Pronoun Processing: Computational, Behavioral, and Psychophysiological Studies in Children and Adults*. Ph.D. dissertation, University of Groningen, Groningen.

van Rij, J., van Rijn, H., and Hendriks, P. (2010). Cognitive architectures and language acquisition: a case study in pronoun comprehension. *J. Child Lang.* 37, 731–766. doi: 10.1017/S0305000909990560

van Rij, J., van Rijn, H., and Hendriks, P. (2013). How WM load influences linguistic processing in adults: a computational model of pronoun interpretation in discourse. *Topics Cogn. Sci.* 5, 564–580. doi: 10.1111/tops.12029

Vogels, J., Krahmer, E., and Maes, A. (2014). How cognitive load influences speakers' choice of referring expressions. *Cogn. Sci.* 39, 1396–1418. doi: 10.1111/cogs.12205

Vogelzang, M., Hendriks, P., and van Rijn, H. (2015). "Processing overt and null subject pronouns in Italian: a cognitive model," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, eds D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings et al. (Austin, TX: Cognitive Science Society).

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128. doi: 10.1016/0010-0277(83)90004-5

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 23 September 2016
doi: 10.3389/fpsyg.2016.01429

Referential Choice: Predictability and Its Limits

Andrej A. Kibrik[1,2]*, *Mariya V. Khudyakova*[3], *Grigory B. Dobrov*[4], *Anastasia Linnik*[5]* and *Dmitrij A. Zalmanov*[2]

[1] *Department of Typology and Areal Linguistics, Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia,* [2] *Department of Theoretical and Applied Linguistics, Lomonosov Moscow State University, Moscow, Russia,* [3] *Neurolinguistics Laboratory, National Research University Higher School of Economics, Moscow, Russia,* [4] *Consultant Plus, Moscow, Russia,* [5] *Linguistics Department, University of Potsdam, Potsdam, Germany*

We report a study of referential choice in discourse production, understood as the choice between various types of referential devices, such as pronouns and full noun phrases. Our goal is to predict referential choice, and to explore to what extent such prediction is possible. Our approach to referential choice includes a cognitively informed theoretical component, corpus analysis, machine learning methods and experimentation with human participants. Machine learning algorithms make use of 25 factors, including referent's properties (such as animacy and protagonism), the distance between a referential expression and its antecedent, the antecedent's syntactic role, and so on. Having found the predictions of our algorithm to coincide with the original almost 90% of the time, we hypothesized that fully accurate prediction is not possible because, in many situations, more than one referential option is available. This hypothesis was supported by an experimental study, in which participants answered questions about either the original text in the corpus, or about a text modified in accordance with the algorithm's prediction. Proportions of correct answers to these questions, as well as participants' rating of the questions' difficulty, suggested that divergences between the algorithm's prediction and the original referential device in the corpus occur overwhelmingly in situations where the referential choice is not categorical.

Keywords: referential choice, non-categoricity, machine learning, cross-methodological approach, discourse production

OPEN ACCESS

Edited by:
Kees van Deemter,
University of Aberdeen, UK

Reviewed by:
Petra Hendriks,
University of Groningen, Netherlands
Chenghua Lin,
University of Aberdeen, UK

Correspondence:
Andrej A. Kibrik
aakibrik@gmail.com
Anastasia Linnik
anastasia.linnik@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: 18 September 2015
Accepted: 06 September 2016
Published: 23 September 2016

Citation:
Kibrik AA, Khudyakova MV, Dobrov GB, Linnik A and Zalmanov DA (2016) Referential Choice: Predictability and Its Limits. Front. Psychol. 7:1429. doi: 10.3389/fpsyg.2016.01429

INTRODUCTION

As we speak or write, we constantly mention various entities, or referents. The process of mentioning referents is conventionally called *reference*. When the speaker's/writer's decision to mention a referent is in place, another discourse phenomenon becomes relevant: *referential choice* that is the process of choosing an appropriate linguistic expression for the referent in question. The question of reference per se, that is of how and why a speaker/writer decides which referent to mention at a given place in discourse, is out of the scope of this paper (cf. the point of Gatt et al., 2014, p. 903, that referential choice is not directly related to the likelihood with which a referent is mentioned), that referential choice is not directly related to the likelihood with which a referent is mentioned). The focus of this study is the phenomenon of referential choice: we explore what guides a speaker/writer in choosing a linguistic expression when s/he has already made a decision to mention a certain referent.

The approach to referential choice adopted in the present study relies on earlier work by Chafe (1976, 1994), Givón (1983), Fox (1987), Tomlin (1987), Ariel (1990), and Gundel et al. (1993). These and other theoretical approaches assumed some kind of a cognitive characterization of a referent that underlies referential choice, such as givenness, topicality, focusing, accessibility, salience, prominence, etc. In terms of the cognitive model developed by Kibrik (1996, 1999, 2011) referential choice is governed by *activation in working memory*. In that model reference per se is claimed to be associated with a distinct cognitive phenomenon of *attention*. Attention and working memory are two related but distinct neurocognitive processes (Cowan, 1995; Awh and Jonides, 2001; Engle and Kane, 2004; Awh et al., 2006; Repovš and Bresjanac, 2006; Shipstead et al., 2015). Accordingly, reference and referential choice, as linguistic manifestations of attention and activation, are related but distinct processes (see Kibrik, 2011, Chap. 10).

As is widely held since Chafe (1976) and Givón (1983), the more given (or salient, accessible) a referent is to the speaker at the moment of reference, the less coding material it requires. In terms of the cognitive model we assume, the main law of referential choice can be formulated as follows:

- If the referent's activation in the speaker's working memory is high, use a reduced referential device. If the referent's activation in the speaker's working memory is low, use a lexically full referential device.

Thus the *basic*, coarse-grained referential choice is between reduced (or attenuated) and lexically full referential devices. In the case of English, it is the distinction between pronouns (personal and possessive), on the one hand, and a variety of full noun phrases, on the other. This distinction is the first level of granularity in the domain of referential options, and all scales and hierarchies that relate givenness (or equivalent concepts) to referential forms (Givón, 1983; Ariel, 1990; Gundel et al., 1993) acknowledge this basic distinction, even though they involve a greater detail in the taxonomy of referential devices. The second level distinction in the domain of referential options is between proper names and descriptions (Anderson and Hastie, 1974; Ariel, 1990; McCoy and Strube, 1999; Poesio, 2000; Heller et al., 2012). There are also further levels of distinction related to varieties of proper names and especially descriptions. In the present study, we mostly concentrate on the first level distinction between pronouns and full noun phrases, and will look briefly into the second level distinction between proper names and descriptions. Our focus is thus different from most work in the current tradition or referring expression generation (REG or GRE, beginning from Dale, 1992 and reviewed in Krahmer and van Deemter, 2012), primarily addressing various types of descriptions. Interestingly, however, Reiter and Dale (2000) recognize that the choice of the "form of referring expressions" (that is, the choice between pronouns, proper names, and descriptions) is the primary one. Krahmer and van Deemter (2012, p. 204) also suggest that first "the form of a reference is predicted, after which the content and realization are determined".

This study is based on a corpus of written English, specifically newspaper (Wall Street Journal) texts. The corpus is annotated in accordance with the MoRA (Moscow Reference Annotation) scheme, detailed in Section "Materials and Methods" below. We assume that written media texts are a good testing ground for our approach. Specific aspects of referential processes differ across various discourse modes and types (see e.g., Fox, 1987; Toole, 1996; Strube and Wolters, 2000; Efimova, 2006; Garrod, 2011), but the basic cognitive principles of referential choice must be shared by all users of a given language and apply to various discourse types.

Example (1) (from the WSJ corpus we explore) illustrates the major referential options.

(1) But beyond this decorative nod to tradition, Ms. Bogart and company head off in a stylistic direction that all but transforms Gorky's naturalistic drama into something akin to, well, farce. The director's attempt to Ø force some Brechtian distance between her actors and their characters frequently backfires with performances that are unduly mannered. Not only do the actors stand outside their characters and Ø make it clear they are at odds with them, but they often literally stand on their heads.

Two referents recur a number of times in (1). They are emphasized with two different kinds of underlining: Ms. Bogart and the actors. The first referent is mentioned with a proper name (title plus last name), a description (the director), as well as with a pronoun (her) and a zero (in an infinitival construction). The second referent is mentioned by two different descriptions (company and actors), pronouns (they, their), and a zero (in a coordinate construction). (In written English, zeroes are not a part of discourse-based referential choice, but they can serve as antecedents; see discussion in Section "Materials and Methods".)

What factors influence actual referential choices in discourse? In usual face-to-face conversation, an entity sometimes become visually available to the interlocutors (via shared attention), and that may be enough for using an exophoric pronoun without any antecedent (see e.g., Cornish, 1999). In written discourse, however, factors affecting referential choice are mostly associated with (i) the referent's internal properties and (ii) the discourse context. Referent's internal properties vary from most inherent, such as animacy, to more fluid, such as being or not being the protagonist of the current discourse. The factors of discourse context are diverse and include the following groups:

- those related to a prospective anaphor, such as the ordinal number of the given mention in the given discourse
- those related to the antecedent's properties, such as its grammatical role (subject, object, etc.)
- those related to discourse structure, such as the distance between the anaphor and the antecedent, measured in the number of clauses or paragraphs.

Referential choice thus belongs to a large family of *multi-factorial processes*, generally characteristic of language production. Most of the factors employed in our study, such as animacy, grammatical role, or distance to antecedent, have been proposed in prior literature, in particular (Paducheva, 1965; Chafe, 1976; Grimes, 1978; Hinds, 1978; Clancy, 1980; Marslen-Wilson et al., 1982; Givón, 1983; Brennan et al., 1987; Fox, 1987; Tomlin, 1987; Ariel, 1990; Gernsbacher, 1990; Gordon et al., 1993; Dahl and Fraurud, 1996; Kameyama, 1999; Yamamoto, 1999; Strube and Wolters, 2000; Arnold, 2001; Stirling, 2001; Tetreault, 2001; Arnold and Griffin, 2007; Kaiser, 2008; Fukumura and van Gompel, 2011, 2015; Fukumura et al., 2013; Fedorova, 2014; Rohde and Kehler, 2014, i.a.). There is no room here to review this literature in detail, but many of these studies are discussed in Kibrik (2011); see also recent reviews in van Deemter et al. (2012) and Gatt et al. (2014). In some of the above-mentioned studies one of the factors was emphasized, while others were ignored or shaded. We find it important to take as many relevant factors as possible into account, as they actually operate in conjunction.

Within the cognitive model we assume, these factors are interpreted as *activation factors*, contributing to the cumulative current referent's activation. This cognitive model of referential choice is depicted in **Figure 1** (see further specification of the model in Sections "Discussion: Referential Choice Is Not Always Categorical" and "Experimental Studies of Referential Variation"). Two kinds of activation factors operate in conjunction and determine a referent's current degree of activation, which in turn predicts referential choice.

In Kibrik (1996, 1999) a simple mathematical model was developed, capturing the multiplicity of factors and their relative contributions to referent activation and, therefore, to the ensuing referential choice. In those studies referent's current activation level was assessed numerically, as a so-called activation score ranging from a minimal to a maximal value. In this paper, in contrast, we present a study based on machine learning techniques, in which we supply activation factors' values to algorithms and obtain predictions of referential choice as an output. Therefore, the activation component remains hidden within the algorithm, and only mappings of activation factors upon referential options are explicit. In this respect this study is similar to most other studies or referential choice cited above, as well as to the studies based on annotated referential corpora, such as Poesio and Artstein (2008) and Belz et al. (2010). Still we find it important to keep the larger picture in mind and recognize that in the human cognitive system referent's activation level mediates between the relevant factors and the actual referential choice.

We pursue two goals in this paper. The first goal is *to predict referential choice* as reliably as possible. We explore a corpus of English written discourse and use machine learning techniques to predict referential choice maximally close to the original texts. This part of the study is reported in Section "Corpus-Based Modeling". In the course of this work it is found that even well-trained algorithms sometimes diverge from the original referential choices in the corpus texts.

That brings us to the second goal of our research: is 100% accurate prediction of referential choice possible in principle? In addressing this question, we consider the possibility that certain instances of divergence between the predicted and original forms may be due to the *incomplete categoricity* of referential choice. In Section "Experimental Studies of Referential Variation", we submit the instances of divergence to an experimental assessment by human participants, in order to see whether people accept referential variation in the spots where divergences take place.

The discussion of our findings and concluding remarks follow in Section "General Discussion".

CORPUS-BASED MODELING

Related Work

During the last twenty years or so a number of corpus resources for studies of coreference and reference production has appeared, including MUC-6/-7 (Chinchor and Sundheim, 1995; Grishman and Sundheim, 1995; Chinchor and Robinson, 1997), the ASGRE challenge (Gatt and Belz, 2008), the GNOME corpus (Poesio, 2000, 2004), the ARRAU corpus (Poesio and Artstein, 2008), and the GREC-08, -09, -10 challenges (Belz and Kow, 2010; Belz et al., 2008, 2009). Among these, the series of studies conducted for the GREC (Generating Referring Expressions in Context) challenges were somewhat similar in their goals to the present study: they predicted the form of a referring expression (common noun, name/description, pronoun, or "empty" reference) in Wikipedia articles about cities, countries, rivers, and people. One of the successful algorithms, a memory-based learner (Krahmer et al., 2008), was able to predict the correct type of referring expression in 76.5% of the cases. Krahmer et al. (2008) used automatic

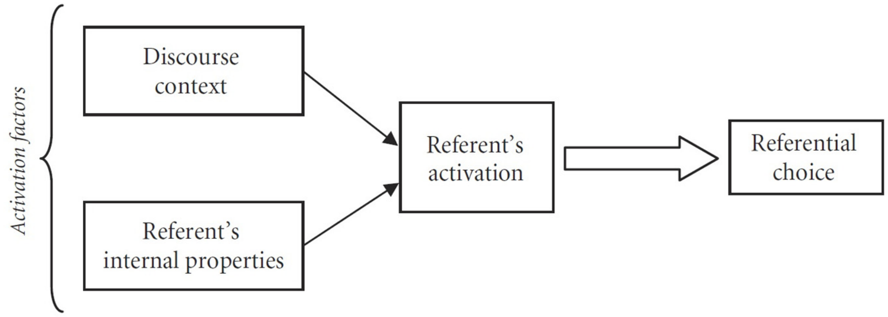


FIGURE 1 | Cognitive model of referential choice (cf. Kibrik, 2011, p. 61).

language processing tools to mark the following parameters for every entity: competition, position in the text, syntactic and semantic category, local context (POS tags), distance to the previous mention in sentences and NPs, main verb of the sentence, and syntactic patterns of three previous mentions. The systems in the 2010 GREC challenge used various sets of factors and machine learning techniques; for example, Greenbacker and McCoy (2009) used such features as competition, parallelism, and recency. The best system's precision in the prediction task reached 82−84%. Zarrieß and Kuhn (2013) report a similarly high prediction accuracy in their study inspired by the GREC tasks on a corpus of German robbery reports. Crucial differences of the present work from the GREC studies are that, first, all referents are considered, not just the main topic referent of each article, and, second, semantic discourse structure is taken into account. Recent reviews providing detailed accounts of corpus-based studies of reference production can be found in Krahmer and van Deemter (2012) and Gatt et al. (2014).

Early modeling studies by Kibrik (1996, 1999) were mentioned in Section "Introduction". Grüning and Kibrik (2005) applied the neural networks method of machine learning to the same small dataset as in Kibrik (1999); that study showed that machine learning is in principle appropriate for modeling multi-factorial referential choice and raised the question of creating a much larger and statistically valid corpus designed for referential studies. Several studies of our group addressed a corpus of Wall Street Journal texts, somewhat larger than the one used in the present paper (Kibrik and Krasavina, 2005; Krasavina, 2006) and used the annotation scheme proposed in Krasavina and Chiarcos (2007). More recently we developed the MoRA (Moscow Reference Annotation) scheme and conducted machine learning studies on the corpus data, looking into the basic referential choice (two-way choice between pronouns and full NPs) and the three-way choice between pronouns, proper names, and descriptions (Kibrik et al., 2010; Loukachevitch et al., 2011). Compared to our previous publications, in the present study we have substantially improved the quality of corpus annotation and modified the annotation scheme and the machine learning methods.

A number of studies emphasized the role of discourse structure in referential choice. In his classical work, Givón (1983) introduced the concept of linear distance from an anaphor back to the antecedent, measured in discourse units such as clauses. Other studies (Hobbs, 1985; Fox, 1987; Kibrik, 1996; Kehler, 2002) underlined the contribution of the semantic structure of discourse, including the hierarchical structure. Several models of discourse-semantic relations have been proposed in the recent decades (see Hobbs, 1985; Polanyi, 1985; Wolf et al., 2003; Miltsakaki et al., 2004; Joshi et al., 2006, i.a.), one of the best known being Rhetorical Structure Theory (RST) (Mann and Thompson, 1987; Taboada and Mann, 2006). RST represents text as a hierarchical structure, in which each node corresponds to an elementary discourse unit (EDU), roughly equaling a clause. Fox (1987) demonstrated a possible connection between reference and RST-based analysis of dicourse, and Kibrik (1996) introduced the measurement of rhetorical distance (RhD) that captures the length of path between an anaphor EDU and the antecedent EDU along the rhetorical graph; see Section "Materials and Methods". In a neural networks-based study (Grüning and Kibrik, 2005) it was also found that RhD was an important factor. Experimental studies of Fedorova et al. (2010b, 2012) demonstrated that RhD is a relevant factor affecting referent activation in working memory, as well as reference resolution in the course of discourse comprehension.

The WSJ MoRA 2015 corpus employed in this paper (we used the name "RefRhet corpus" for earlier versions in previous publications) is based on a subset of texts of the RST Discourse Treebank, developed by Daniel Marcu and his collaborators (Carlson et al., 2002). This allows us to combine our own annotation (see Materials and Methodsith the rhetorical annotation produced by the Marcu's team, and to compute RhD on the basis of their annotation. To the best of our knowledge, corpora intended for referential studies and containing discourse semantic structure annotation are few on the market; cf. the German corpus Stede and Neumann (2014). An English language resource comparable to ours in using discourse semantic structure as a part of referential annotation is the so-called C-3 corpus outlined in Nicolae et al. (2010). As these authors correctly state,

> "the most widely known coreference corpora < … > are annotated with relations between entities, not between discourse segments. The most widely known coherence corpora are Discourse GraphBank (Wolf & al., 2003), RST Treebank (Carlson & al., 2002), and Penn Discourse Treebank (Prasad & al., 2008), none of which was annotated with coreference information." (Nicolae et al., 2010, p. 136).

Nicolae et al.'s (2010) project is similar to ours in that they picked an already existing corpus annotated for discourse semantic relations and added further annotation for the purposes of modeling reference. Unlike us, however, they chose not the RST Discourse Treebank but the Discourse GraphBank of Wolf et al. (2003). The latter corpus is based on a less constrained kind of discourse representation compared to RST; see discussion in Marcu (2003), Wolf et al. (2003), and Wolf and Gibson (2003).

Referential annotation added by Nicolae et al. (2010) includes primarily types of entities (persons, organizations, locations, etc.), referential status (specific, generic, etc.) and referential form (pronoun, proper name, description, etc.). The number of entity types is greater than in our annotation scheme, but in general there are much fewer parameters involved. In particular, it seems that the syntactic role of anaphors and antecedents is not annotated. Generally Nicolae et al. (2010) followed the ACE (Automatic Content Extraction, 2004) guidelines principles of coreference annotation. They developed their own annotation tool. We are not aware of specific modeling studies based on the C-3 corpus.

A variety of algorithms have been used in computational studies of referential choice. One of the well-known early algorithms is the so-called incremental algorithm that was used by Dale and Reiter (1995) to predict the choice of attributes in descriptions. Modifications of this algorithm include the ones developed by van Deemter (2002) and Siddharthan and Copestake (2004), i.a.. In the 2000s, with the development of corpora for referential studies, researchers began to use classical machine learning algorithms and methodology to analyze some features of referential expressions. For example, in Cheng et al. (2001) the classification task was to determine the NP type, and the corpus annotation was used to train a classifier. The authors used the CART (Classification and Regression Trees) classifier and achieved 67 and 75% accuracy on different text sets by cross-validation procedure. Early corpus- and machine learning-based studies similar to ours in design are Poesio et al. (1999) and Poesio (2000). In the studies related to the GREC challenges (Belz and Varges, 2007; Belz and Kow, 2010), the algorithms had to identify the correct referring expression from a provided set. Participants used various methods and features to perform the task. For example, in 2008 they were: Conditional Random Fields with a set of features encoding the attributes given in the corpus, information about intervening references to other entities, etc. (UMUS system); a set of decision tree classifiers that checked the length of referring expressions and correctness of pronouns (UDEL system); XRCE system that used a great number of features with levels of activation. Other studies applying machine learning specifically to discourse reference include Jordan and Walker (2005), Viethen et al. (2011), and Ferreira et al. (2016). Also, there is a number of studies in which machine learning was used in other language generation tasks, such as prediction of adjective ordering (Malouf, 2000), content selection (Kelly et al., 2009), accent placement (Hirschberg, 1993), sentence planning (Walker et al., 2002), automated generation of multi-sentence texts (Hovy, 1993), as well as other tasks (e.g., Dethlefs and Cuayáhuitl, 2011; Dethlefs, 2014; Stent and Bangalore, 2014).

Materials and Methods

The Corpus

The WSJ MoRA 2015 corpus explored in this study consists of Wall Street Journal articles from the late 1980s, including broadcast news, analytical reviews, cultural reviews, and some other genres. Text length varies from 70 words to about 2000 words, the average length being 375 words. A general quantitative characterization of the WSJ MoRA 2015 corpus appears in **Table 1**.

Referential annotation of the corpus consists of two parts: annotation of referential devices and annotation of candidate activation factors. We consider these two kinds of annotation in turn.

TABLE 1 | The WSJ MoRA 2015 corpus: a quantitative characterization.

Feature	Comment	Number in corpus
Texts		64
Paragraphs		511
Sentences		976
Elementary discourse units (EDU)	EDU segmentation of texts is automatically extracted from the RST Discourse Treebank	2928
Words		23952

Annotation of referential devices

Referential devices are technically named markables that is those referential expressions that can potentially corefer. Coreferential expressions form a *referential chain*. Non-first members of a referential chain are termed *anaphors* below. The breakdown of markables by type is shown in **Table 2**.

Note that not every markable in the corpus is actually used for analysis. First, there are 2580 singleton markables that are not linked to any other markable by a coreference relation and are not

TABLE 2 | Types and numbers of markables (referential expressions).

	Type of markable	Comment	Number in corpus
1.	**Reduced referential devices**	Sum of #2 to #7	**1373**
2.	Personal pronouns		495
3.	Possessive pronouns		264
4.	Zeroes		375
5.	Demonstratives		67
6.	Relative pronouns		135
7.	Other		37
8.	**Full noun phrases**	Sum of #9 and #18 minus #27*)	**5042**
9.	**Descriptions**	Sum of #10 to #15	**3517**
10.	The-descriptions		1241
11.	A-descriptions		420
12.	Bare descriptions		1200
13.	Demonstrative descriptions	E.g. *this house*	88
14.	Possessive descriptions	E.g. *his house, the company's shares*	490
15.	Other		78
	Special subtypes		
16.	Attributive descriptions	E.g. *the American president; the first American president who was elected...*	1458
17.	Numeral descriptions	E.g. *the two books*	136
18.	**Proper names**	Sum of #19 to #25*)	**1681**
19.	First names		21
20.	Last names		229
21.	First plus last names		193
22.	Initials plus last names	E.g. *G.W.Bush*	1
23.	Non-persons	Names of countries, organizations, units, etc.	915
24.	Acronyms	E.g. *GE, the US*	277
25.	Other		45
	Special subtype		
26.	Titled proper names	E.g. *Mr. Bush*	162
27.	**Mix: description plus proper name**	E.g. *President Bush*	**156**
	TOTAL		**6415**

**Special subtypes in lines 16–17 and 26 cross-cut the mutually exclusive subtypes appearing in lines 10–15 and 19–25, respectively, and therefore are not summed with those in the counts shown in lines 9 and 18.*

pertinent to referential choice. (They are nevertheless annotated, as they are taken into account when the values for the factor "distance in markables" are calculated.) In the modeling task we only use those markables that form referential chains. Second, certain types of referential expressions are only considered as antecedents, but not as anaphors in our analysis of referential choice. This concerns the following categories:

– indefinite descriptions (introduced by indefinite determiners, such as *a(n), some, few,* etc.);
– bare descriptions;
– all types of pronouns other than personal and possessive;
– first and second person pronouns;
– zero references.

In particular, quite common zero references in English only appear in fixed syntactic contexts, such as coordinate, gerundial, and infinitival constructions; at least this applies to the kind of written English we explore (cf. Scott, 2013). Syntactically induced zeroes should not be treated as a discourse-based referential option on a par with third person pronouns and full NPs. At the same time, zeroes make bona fide antecedents, so they must be annotated as markables in a referential corpus[1]. Similar reasoning applies to relative pronouns. In written discourse, nominal demonstratives such as *that* typically refer to situations rather than entities.

In the corpus, there are 777 referential chains that comprise at least one anaphor, meeting the above-listed requirements (i.e., is not a bare description, a zero, etc.). Such chains include 3199 markables used in the modeling tasks. Average chain length is 4.1 markables, and the maximum length of a chain is 52 markables.

We thus address the basic referential choice between third person personal/possessive pronouns and full noun phrases. **Table 3** shows the numbers of anaphors in the corpus.

Annotation of candidate activation factors

The second part or referential annotation addresses candidate activation factors that is parameters that are potentially useful for the prediction of referential choice. The complete list of candidate factors used in this study is shown in **Table 4**. For each factor, its values included in the study are listed after a colon. Most of the factors' values are derived from the MoRA scheme annotation, but some are computed automatically.

In **Table 4**, the factors are listed in four groups. In the terms of **Figure 1**, the group 1 factors roughly correspond to the "Referent's internal properties" activation factors, while group 2–4 factors to the "Discourse context" activation factors. For the sake of brevity, the logic of factors is somewhat simplified in **Table 4**. In particular, most factors include the value "other" that we omit here. Several of the factors call for clarifying comments.

Protagonism means referent's centrality in discourse. Two models of protagonism were used (Linnik and Dobrov, 2011): in the first one, to each referent corresponds the ratio of its referential chain length to the maximal length of a referential chain in the text; in the second model, to each referent corresponds the ratio of its chain length to the gross number of markables in the text. In both instances, the most frequently mentioned referent is the same, but relative weights of referents may be different.

Regarding the "Type of phrase" factor, it is important to explain why we consider prepositional phrases (such as *of the president* or *with her*) a particular type of phrase, rather than a combination of a preposition with a referential device (noun phrase). First, referential choice may depend on whether the antecedent or the anaphor is a plain noun phrase, or a noun phrase subordinate to a preposition (that is, constitutes a prepositional phrase); so this information must be retained. Second, consider English *'s*- and *of*-genitives. The former are

TABLE 3 | Anaphor types.

Anaphor type	Number used for analysis
Third person pronouns (personal or possessive)	585 (26.0%)
Descriptions	856 (38.1%)
Proper names	807 (35.9%)
Total	2248 (100%)

TABLE 4 | Candidate factors of referential choice.

(1) Referent's factors

- Animacy: animate, inanimate, collective (*for such entities as organizations*)
- Gender (for animate referents only): masculine, feminine, mixed (*for groups of people with various or unspecified gender*)
- Person: 1, 2, 3
- Number: singular, plural
- Protagonism: *numeric value*

(2) Anaphor's factors

- Ordinal number of referent mention in the referential chain: *integer*
- Type of phrase: noun phrase, prepositional phrase
- Grammatical role: subject, direct object, indirect object, oblique (with preposition), attribute, *'s*-genitive, *of*-genitive, postpositive specification

(3) Antecedent's factors

- Type of phrase (values same as in the section "Anaphor's factors")
- Grammatical role (values same as in the section "Anaphor's factors")
- Referential form:
  - ○ pronoun: personal, possessive, demonstrative, relative, zero
  - ○ description: a-description, the-description, bare description, demonstrative description, possessive description
    - ○ attributive
    - ○ numeral
  - ○ proper name: first, last, first and last, initials and last, non-person, acronym
  - ○ Antecedent length, in words: *integer*

(4) Distances between anaphor and antecedent

- Distance in words: *integer*
- Distance in all markables: *integer*
- Number of markables in chain from the anaphor back to the nearest full NP antecedent: *integer*
- Linear distance in EDUs: *integer*
- Rhetorical distance (RhD) in elementary discourse units: *integer*
- Distance in sentences: *integer*
- Distance in paragraphs: *integer*

[1] No zero symbols are introduced into the corpus for the purposes of annotation. Instead, we annotate reference on a verb form of which a zero is the subject; cf. this kind of annotation on *to force* and *sprawling*, as shown in **Figure 3**.

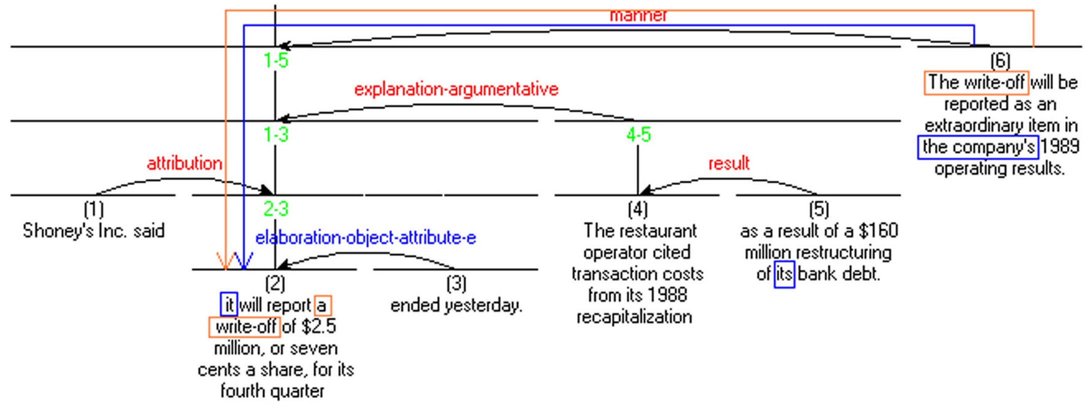

manner
1-5
(6)
The write-off will be reported as an extraordinary item in the company's 1989 operating results.
explanation-argumentative
1-3
4-5
attribution
result
2-3
(1)
Shoney's Inc. said
elaboration-object-attribute-e
(4)
The restaurant operator cited transaction costs from its 1988 recapitalization
(5)
as a result of a $160 million restructuring of its bank debt.
(2)
it will report a write-off of $2.5 million, or seven cents a share, for its fourth quarter
(3)
ended yesterday.

FIGURE 2 | Example of a rhetorical graph from RST Discourse Treebank with examples of RhD computation. The referent 'the write-off' is mentioned in units #2 and #6. Linear distance from #6 back to #2 equals 4. Rhetorical distance (RhD) from #6 to #2 is just 1, as these two nodes are immediately connected to each other in the RST graph, and one only needs one horizontal step along the graph to reach #2. The anaphor *the company* found in unit #6 has its closest linear antecedent in unit #5. However, its closest rhetorical antecedent is again found in #2, directly connected to the anaphor unit #6. Arrows indicate paths along the RST graph one needs to travel to reach an antecedent.

inflectional word forms and cannot be divided into a referential device and a separate unit, and it is reasonable to treat the two different kinds of genitives in the same way. More generally, in many languages, equivalents of English prepositions would be case endings, and nobody would deduct these from referential expressions.

Most of the distance factors are identifed for the closest linear antecedent. In contrast, RhD is computed from the anaphor back to the nearest rhetorical antecedent along the hierarchical graph. **Figure 2** presents an example of the RST Discourse Treebank annotation, as well as illustrates the difference between the linear and the rhetorical antecedents, and the corresponding distances. Principles of RhD computation were outlined in Kibrik and Krasavina (2005).

In all, 25 potentially relevant activation factors are extractable from the annotated WSJ MoRA 2015 corpus; these are independent variables in the computational models discussed below. The parameter *anaphor's referential form* is the predicted, or dependent, variable.

Each text of the WSJ MoRA 2015 corpus was annotated by two different annotators, and each pair of annotations was compared with the help of a special script that identified divergences. All problematic points were fixed by an expert annotator. The corpus was subsequently cross-checked with a variety of techniques and corrected by the members of our team.

Figure 3 provides a screenshot from the MMAX2 annotation tool (Müller and Strube, 2006) for the same text excerpt that was used as Example (1) in Section "Introduction". Here, all expressions that refer to "Ms. Bogart" are highlighted and grouped into one referential chain with lines that mark coreference.

A special property of the MoRA scheme is the annotation of *groups*. A group is a set of markables that, collectively, serve as an antecedent of an anaphor. In **Figure 3**, two groups are present, marked with curly brackets and with italics: {[*Ms. Bogart*] *and* [*company*]} and {*between* [[*her*] *actors*] *and* [[*their*] *characters*]}. Later on in the text, there is indeed the markable [of the ensemble], the antecedent of which is {[*Ms. Bogart*] *and* [*company*]}.

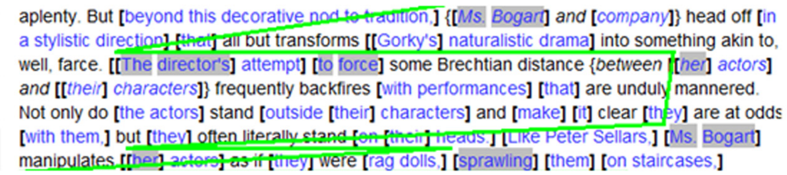
Computational Modeling

In this study we use the system Weka[2] (see Hall et al., 2009) that includes many algorithms of machine learning, as well as automated means of algorithms' evaluation. Several types of algorithms, or classifiers, are used. We consider the wide variety of used algorithms as an important methodological property of our study, distinguishing it from most other studies in reference production.

First, we use a logical algorithm (decision trees C4.5) as it lends itself to natural interpretation. Second, we use logistic regression because its results often exceed those of logical algorithms in quality. In addition, we use the so-called classifier compositions: bagging (Breiman, 1996) and boosting (Freund and Schapire, 1996). These composition algorithms use, as a source of their parameters, another machine learning algorithm that we will call the base algorithm. Using the base algorithm, composition algorithms construct multiple models and combine their results. As was shown in several experimental studies (for example, Schapire, 2003), composition algorithms or their modifications "performed as well or significantly better than the other methods tested" (Schapire, 2003, p. 162).

In the boosting algorithm the base algorithm undergoes optimization. An adaptation of classifiers is performed, that is, each additional classifier applies to the objects that were not properly classified by the already constructed composition. After each call of the algorithm the distribution of weights is updated. (These are weights corresponding to the importance of the training set objects.) At each iteration the weights of each wrongly classified object increase, so that the new classifier focuses on such objects. Among the boosting algorithms,

[2] http://www.cs.waikato.ac.nz/ml/weka



aplenty. But [beyond this decorative nod to tradition,] {[Ms. Bogart] *and* [*company*]} head off [in a stylistic direction] [that] all but transforms [[Gorky's] naturalistic drama] into something akin to, well, farce. [[The director's] attempt] [to force] some Brechtian distance {*between* [[*her*] *actors*] *and* [[*their*] *characters*]} frequently backfires [with performances] [that] are unduly mannered. Not only do [the actors] stand [outside [their] characters] and [make] [it] clear [they] are at odds [with them,] but [they] often literally stand [on [their] heads.] [Like Peter Sellars,] [Ms. Bogart] manipulates [[her] actors] as if [they] were [rag dolls,] [sprawling] [them] [on staircases.]

FIGURE 3 | A sample of text annotation in MMAX2.

AdaBoost was used in our modeling with C4.5 as the base algorithm.

Bagging (from "bootstrap aggregating") algorithms are also algorithms of composition construction. Whereas in boosting each algorithm is trained on one and the same sample with different object weights, bagging randomly selects a subset of the training samples in order to train the base algorithm. So we get a set of algorithms built on different, even though potentially intersecting, training subsamples. A decision on classification is made through a voting procedure in which all the constructed classifiers take part. In the case of bagging the base algorithm was also C4.5.

In order to control the quality of classification, the cross-validation procedure was used:

(1) The training set is divided into ten parts.
(2) A classifier operates on the basis of nine parts.
(3) The constructed decision function is tested on the remaining part.

The procedure is repeated for all possible partitions, and the results are subsequently averaged. The criterion for choosing both an optimal set of features and an algorithm is *accuracy* that is the ratio of properly predicted referential expressions to the overall amount of referential expressions. As was pointed out above, all the independent variables contained in **Table 4** were treated as candidate factors of referential choice and included into our machine learning studies.

Results

Predicting Basic Referential Choice

The results of modeling the basic choice between reduced and full referential devices are given in **Table 5**. The baseline means the frequency of the most frequent referential option, that is, full noun phrase. If an algorithm always predicted the most frequent option, its accuracy would equal that option's frequency. **Table 5** also includes information on three additional measures assessing the quality of classification: precision, recall, and F1 (or harmonic mean).

The results yielded by any of the algorithms surpass the baseline substantially. At the same time, with the given set of factors all the algorithms demonstrate very close results; in particular, the accuracy rate is in the vicinity of 89—90%. The boosting algorithm fairs somewhat better than the others, but its difference from the other algorithms is not statistically significant. (We performed the McNemar's test of statistical significance, in accordance with the method described in Salzberg, 1997.)

The confusion matrix (i.e., information on the amount of divergent predictions done by a classifier) for the boosting algorithm appears in **Table 6**. The model predicts over 93% of full NPs correctly, but is less effective with respect to pronouns: only 77% are predicted correctly. Such difference in performance can be explained by the class imbalance in the task: machine learning algorithms "prefer" to predict the most frequent class (full NP in our case) and thus achieve higher overall accuracy (Longadge et al., 2013). It is hardly possible to avoid class imbalance in a corpus-based study, in which relative frequencies of tokens consitute an inherent part of the data.

Interpreting Decision Trees

Among the machine learning algorithms, decision trees may be particularly telling in explicitly specifying the concrete role of certain factors. For our corpus, a decision tree was generated that comprised 110 terminal nodes each corresponding to a specific prediction rule. Consider the following branch from the decision tree: if the anaphor is a prepositional phrase and its antecedent lies within the same sentence, then it is most probable that a full noun phrase will be chosen, not a pronoun. Of 100 instances observed, only 8 display pronominalization. A typical example can be seen in (2).

TABLE 5 | Prediction of the basic referential choice.

Algorithm	Accuracy	Full NP			Pronoun		
		Precision	Recall	F1	Precision	Recall	F1
Baseline	74.0%	74.0%	1	85.0%	0	0	0
C4.5 algorithm	88.9%	91.7%	92.0%	91.9%	77.3%	76.7%	77.0%
Logistic regression	88.6%	91.5%	92.6%	92.1%	78.5%	76.0%	77.2%
Bagging	89.4%	91.9%	93.6%	92.7%	81.0%	76.8%	78.9%
Boosting	89.8%	92.2%	93.6%	92.8%	80.9%	77.4%	79.1%

TABLE 6 | Confusion matrix for the boosting algorithm, basic referential choice.

	Predicted full NP	Predicted third person pronoun	Total
Original full NP	1556 (93.6%)	107 (6.4%)	1663 (100%)
Original pronoun	132 (22.6%)	453 (77.4%)	585 (100%)

(2) Israel has launched a new effort to prove the Palestine Liberation Organization continues to Ø practice terrorism, and thus to persuade the U. S. to break off talks with the group.

This finding is quite surprising, given the closeness of the anaphor to the antecedent. The specific explanation of the finding is yet to be determined, but it is clear that the decision tree algorithms provide a source of new cause-effect generalizations about referential choice that would otherwise remain unrevealed.

Factors' Contribution

What is the role of individual factors to the success of prediction? In order to evaluate such role, we have applied the boosting algorithm to different subsets of factors in order to find out the individual contribution of factors or their combinations. The results are provided in **Table 7**.

We used a number of distance measurements in this study. The data in **Table 7** suggests that this group of factors is essential for successful prediction. As the distance factors are highly correlated, using any of them increases accuracy dramatically. Accuracy increases further if two or three distance factors are included. The non-distance factors have complex impact on accuracy: eliminating them one by one does not impair prediction significantly, but removing all of them results in a significant decrease of accuracy and is therefore inadvisable.

An earlier study of our group (Loukachevitch et al., 2011) specifically looked into the selection of factors and explored the relationships between them. Models based on various subsets of the factors were tested, and it was demonstrated that none of those models surpassed the full set of factors in classification quality. Deduction of each individual factor led to some deterioration of prediction. This makes us believe that the full set of factors used in our studies can hardly be reduced without detriment to the quality of prediction.

TABLE 7 | The significance of factors in modeling the basic referential choice (boosting with 50 iterations).

Factors	Accuracy(%)
All factors	89.8
— without animacy	89.4
— without protagonism	89.7
— without the anaphor's grammatical role	88.3
— without the antecedent's grammatical role	89.2
— without grammatical role	87.7
— without the antecedent's referential form	89.4
All non-distance factors only	75.5
— plus distance in all markables	82.5
— plus distances in words and paragraphs	87.2
— plus RhD, distance in words, and distance in sentences	88.7
All distance factors only	83.2

Modeling the Three-way Referential Choice

The set of candidate activation factors employed in this study is derived from the vast tradition of studies on basic referential choice. We have reached a significant success in predicting the basic choice. Now, what governs the second-order choice between the types of full noun phrases, that is, proper names and descriptions? Studies of these issues are relatively few (cf. Anderson and Hastie, 1974; Arutjunova, 1977; Seleznev, 1987; Ariel, 1990; Vieira and Poesio, 1999; Enfield and Stivers, 2007; Helmbrecht, 2009; Heller et al., 2012). We have experimentally applied our set of factors to the three-way choice between third person pronouns, proper names, and descriptions. The results can be seen in **Table 8**. The baseline is the frequency of descriptions, the most frequent referential option.

The fairly high accuracy of prediction we have obtained for the three-way task is intriguing. Apparently, the factors responsible for the choice between proper names and descriptions substantially intersect with our basic set of factors. This issue requires further investigation.

Note that in the three-way task boosting again demonstrates the highest results, as it did in the two-way task. Even though the advantage of boosting over the other methods again is not statistically significant, the tendency of its good performance motivates our solution to employ this method in the subsequent part of this study. (However, if we used another algorithm, at least one of those included in our study, the difference would be minimal.)

Discussion: Referential Choice Is Not Always Categorical

Even though the machine learning modeling was quite successful, the accuracy of prediction of the basic referential choice is still quite away from 100%. An important question arises: if we continue improving our annotation (e.g., by extending the set of factors) and tuning up the modeling procedure, can referential choice be ultimately predicted with the accuracy approaching 100%? In other words, is the 10% difference between the algorithm's prediction and the original texts due to certain shortcomings of our methods or to some more fundamental causes? We propose that complete accuracy may not be attainable due to the nature of the process of referential choice.

Referential choice appears to not be a fully categorical and deterministic process. True, there are many instances in which

TABLE 8 | Prediction of the three-way referential choice.

Algorithm	Accuracy (%)
Baseline	38.1
C4.5 Decision tree algorithm	72.3
Logistic regression	73.5
Bagging	73.1
Boosting	75.7

only a pronoun or only a full noun phrase is appropriate, but there are also numerous instances in which more than one referential option can be used. This issue was explored in Kibrik (1999, p. 39), and the basic referential choice was represented as a scale comprising five potential situations:

(3) i. full NP only
ii. full NP, ?pronoun
iii. either full NP or pronoun
iv. pronoun, ?full NP
v. pronoun only.

In (3), situations i and v are fully confident, or categorical, in the sense that language speakers would only use this particular device at the given point in discourse. Situations ii and iv suggest that, in addition to a preferred device, one can marginally use an alternative question-marked device. Finally, situation iii means free variation. In Kibrik (1999) specific referent mentions were attributed to five categories via an experimental procedure. Participants were offered modified versions of the original text, in which referential options were altered – for example, a full noun phrase was replaced by a pronoun or vice versa. Participants were asked to pinpoint infelicitous elements in the text and edit them. As a result of this procedure, some referential devices were assessed as categorical (types i, v). Other referential devices were judged partly (types ii, iv) or fully (type iii) alterable, or non-categorical. (Refer to the original publication for further details.) From the cognitive perspective, this can be interpreted as a mapping from the continuous referent activation to the binary formal distinction, as shown in **Figure 4**. That is, the formulation of the main law of referential choice, as offered in Section 1, suggests an overly categorical representation. It only captures correctly the two poles of the activation scale, but there are intermediate grades of activation in between that lead to less than categorical referential choice. The model of referential choice that we propose, as shown in **Figure 4**, differs from the well-known hierarchies of Givón (1983), Ariel (1990), and Gundel et al. (1993) in two respects. First, it explicitly recognizes a continuous cognitive variable, and second, it only focuses on the highest level distinction between full and reduced referential devices.

Non-categorical and/or probabilistic nature of referential choice has previously been addressed in a number of studies (e.g., Viethen and Dale, 2006a,b; Belz and Varges, 2007; Gundel et al., 2012; Khan et al., 2012; van Deemter et al., 2012; Engonopoulos and Koller, 2014; Ferreira et al., 2016; Hendriks, 2016; Zarrieß, 2016). For example, the well-known scale of Gundel et al. (1993) is implicational in its nature, and that is a way to partly account for the incomplete categoricity of referential choice. Krahmer and van Deemter (2012), noting that the deterministic approach dominates the field, discuss the studies by Di Fabbrizio et al. (2008) and Dale and Viethen (2010) that proposed probabilistic models accounting for individual differences between speakers. van Deemter et al. (2012, p. 18) remark that the probabilistic approach can be extended to a within-individual analysis:

> Closer examination of the data of individual participants of almost any study reveals that their responses vary substantially, even within a single experimental condition. For example, we examined the data of Fukumura and van Gompel (2010), who conducted experiments that investigated the choice between a pronoun and a name for referring to a previously mentioned discourse entity. The clear majority (79%) of participants in their two main experiments behaved non-deterministically, that is, they produced more than one type of referring expression (i.e., both a pronoun and a name) in at least one of the conditions.

Overall, there is accumulating evidence suggesting that human referential choice is not fully categorical. There are certain conditions in which more than one referential option is appropriate and, in fact, each one would fare well enough. Under such conditions human language users may act differently on different occasions. If so, an efficient algorithm imitating human behavior may legitimately perform referential choice in different ways, sometimes coinciding with the original text and sometimes diverging from it. Therefore, ideal prediction of referential choice should not be possible in principle.

We have designed an experiment in which we attempt to differentiate between the two kinds of the algorithm's divergences from the original referential choices. Of course, there may be instances due to plain error. But apart from that, there may be other instances associated with the inherently non-categorical nature of referential choice.

EXPERIMENTAL STUDIES OF REFERENTIAL VARIATION

Related Work

As was discussed in Section "Discussion: Referential Choice Is Not Always Categorical", referential variation and non-categoricity is clearly gaining attention in the modern linguistic, computational, and psycholinguistic literature. Referential

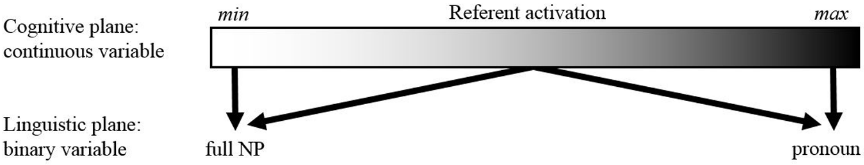


FIGURE 4 | Categorical and non-categorical referential choice.

variation may be due to the interlocutors' perspective taking and their efforts to coordinate cognitive processes, see e.g., Koolen et al. (2011), Heller et al. (2012), and Baumann et al. (2014). A number of corpus-based studies and psycholinguistic studies explored various factors involved in the phenomenon of overspecification, occurring regularly in natural language (e.g., Kaiser et al., 2011; Hendriks, 2014; Vogels et al., 2014; Fukumura and van Gompel, 2015). Kibrik (2011, pp. 56–60) proposed to differentiate between three kinds of speaker's referential strategies, differing in the extent to which the speaker takes the addressee's actual cognitive state into account: egocentric, optimal, and overprotective. There is a series of recent studies addressing other aspects of referential variation, e.g., as a function of individual differences (Nieuwland and van Berkum, 2006), depending on age (Hughes and Allen, 2013; Hendriks et al., 2014) or gender (Arnold, 2015), under high cognitive load (van Rij et al., 2011; Vogels et al., 2014) and even under the left prefrontal cortex stimulation (Arnold et al., 2014). These studies, both on production and on comprehension of referential expressions, open up a whole new field in the exploration of reference.

We discuss a more general kind of referential variation, probably associated with the intermediate level of referent activation. This kind of variation may occur in any discourse type. In order to test the non-categorical character of referential choice we previously conducted two experiments, based on the materials of our text corpus. Both of these experiments were somewhat similar to the experiment from Kibrik (1999), described in Section "Discussion: Referential Choice Is Not Always Categorical" above.

In a *comprehension experiment*, Khudyakova (2012) tested the human ability to understand texts, in which the predicted referential device diverged from the original text. Nine texts from the corpus were randomly selected, such that they contained a predicted pronoun instead of an original full NP; text length did not exceed 250 words. In addition to the nine original texts, nine modified texts were created in which the original referential device (proper name) was replaced by the one predicted by the algorithm (pronoun). Two experimental lists were formed, each containing nine texts (four texts in an original version and five in a modified one, or vice versa), so that original and modified texts alternated between the two lists.

The experiment was run online on Virtual Experiments platform[3] with 60 participants with the expert level command of English. Each participant was asked to read all the nine texts one at a time, and answer a set of three questions after each text. Each text appeared in full on the screen, and disappeared when the participant was presented with three multiple-choice questions about referents in the text, beginning with a WH-word. Two of those were control questions, related to referents that did not create divergences. The third question was experimental: it concerned the referent in point, that is the one that was predicted by the algorithm differently from the original text. Questions were presented in a random order. Each participant thus answered 18 control questions and nine experimental questions. In the alleged instances of non-categorical referential choice, allowing both a full NP and a pronoun, experimental questions to proper names (original) and to pronouns (predicted) were expected to be answered with a comparable level of accuracy.

The accuracy of the answers to the experimental questions to proper names, as well as to the control questions, was found to be 84%. In seven out of nine texts, experimental questions to pronouns were answered with the comparable accuracy of 80%. We propose that in these seven instances we deal exactly with non-categorical referential choice, probably associated with an intermediate level of referent activation. Two remaining instances may result from the algorithms' errors.

The processes of discourse production and comprehension are related but distinct, so we also conducted an *editing experiment* (Khudyakova et al., 2014), imitating referential choice as performed by a language speaker/writer. In the editing experiment, 47 participants with the expert level command of English were asked to read several texts from the corpus and choose all possible referential options for a referent at a certain point in discourse. Twenty seven texts from the corpus were selected for that study. The texts contained 31 critical points, in which the choice of the algorithm diverged from the one in the original text. At each critical point, as well as at two other points per text (control points), a choice was offered between a description, a proper name (where appropriate), and a pronoun. Both critical and control points did not include syntactically determined pronouns. The participants edited from 5 to 9 texts each, depending on the texts' length. The task was to choose all appropriate options (possibly more than one). We found that in all texts at least two referential options were proposed for each point in question, both critical and control ones.

The experiments on comprehension and editing demonstrated the variability of referential choice characteristic of the corpus texts. However, a methodological problem with these experiments was associated with the fact that each predicted referential expression was treated independently, whereas in real language use each referential expression depends on the previous context and creates a context for the subsequent referential expressions in the chain. In order to create texts that are more amenable to human evaluation, in the present study we introduce a flexible prediction script.

Human Evaluation

Preparation of Experimental Material: Flexible Prediction

The modeling method presented in Section "Corpus-Based Modeling" predicts referential choice at each point in discourse where a referential expression is found in the original text. For each referent, if the predicted choice at point *n* diverges from the original one, the subsequent referential expression *n+1* is again predicted by the algorithm on the basis of the original antecedent, rather than on the basis of the previous prediction. This is a traditional and valid method to generally evaluate the accuracy of the algorithm's operation; however, in an experimental setting,

[3] https://virtualexs.ru

where a human evaluation of the whole text is involved, such method is problematic. In order to make referential choices more natural, it is desirable to create a new version of a referential chain, such that a prediction at point *n+1* takes into account what the algorithm had predicted at point *n*.

For human evaluation, we have created a flexible modeling script. The selected referential chain is excluded from the data used for machine learning, so that the training data is kept separate from the test data. The boosting algorithm is run for each member of the chain. If there is a discrepancy between the algorithm's choice and the original choice, it is the predicted referential expression that is used as the antecedent for the subsequent prediction. In this approach each instance of referential choice depends on all previous choices, which is more realistic from the cognitive point of view. We have made changes to the original texts according to the boosting predictions, so that new modified texts were created for each of the two evaluation studies: expert evaluation and experimental evaluation.

Two Stages of Human Evaluation

Human evaluation of predicted referential expressions was performed in two stages. The first stage is a rough evaluation of all the divergences of predicted referential forms from the original texts, done by a single expert. The goal of expert evaluation is to outline a distinction between crude algorithm's errors, leading to a linguistic ill-formedness or a change in the original meaning of a text, and those divergences that may be actually acceptable for a human language user.

The second stage of human evaluation is an experiment with native speakers of English. In contrast to expert evaluation, at the stage of experimental evaluation we select a subset of divergences and present those to multiple participants.

Expert Evaluation

Materials and Methods

Out of the 64 corpus texts, 48 texts demonstrated divergences from the original ones. These texts contained the total of 229 instances of divergence, including 95 predicted pronouns (instead of original full NPs) and 134 predicted full NPs (instead of original pronouns). For the purpose of expert evaluation modified versions of all 48 texts were created, with the use of the flexible script. In the modified texts, original full NPs were replaced by pronouns (with the proper number, gender, and case features), and, conversely, original pronouns were replaced by the most obvious descriptive designation of the referent (same as used in the text elsewhere), such as *the company* for 'General Electric' or *the president* for 'George Bush'.

The modified texts were analyzed by one of the coauthors of this paper. As a result of text assessment, the following most common types of undoubted referential errors were detected: use of a full NP in the context of syntactic anaphora and non-cataphoric third person pronouns at the beginning of a referential chain. Example (4) demonstrates a text excerpt with two predictions not matching the referential expressions found in the original text. The original referential expressions are underlined, and the divergent predictions of the algorithm are indicated in brackets, followed by a specific referential form as used in the experiment. Prediction < 2 > was rated by the expert as potentially fitting, whereas prediction < 1 > was rated as an obvious error, namely a full NP predicted in the context of syntactic anaphora.

(4) Like Brecht, and indeed Ezra Pound, Ms. Bogart has said that < 1 > her [full NP: *the director's*] intent in such manipulative staging of the classics is simply an attempt to make it new. Indeed, during a recent post-production audience discussion, < 2 > the director [pronoun: *she*] explained that her fondest artistic wish was to find a way to play < . . . >

Results

The analysis detected 26 undoubted referential errors that constituted 11% of all divergent predictions and just 1.2% of all referential choices predicted by the algorithm (that is, of 2248 anaphors, see **Table 3**).

Results of expert evaluation suggest that, from a reader's point of view, replacement of an original referential expression by the predicted one mostly does not lead to an obvious referential error. In the texts analyzed, the traditionally measured accuracy of prediction was 90%; however, it appears that, out of the remaining 10%, there were only 1.2% of instances in which a predicted referential expression was rated as completely inappropriate. We interpret this finding as follows: it is not all of divergences of algorithm's prediction from the original texts that are due to error, and the traditional approach to measuring the accuracy of prediction may conceal the difference between the natural variability of referential choice and inaccurate algorithm performance.

Experimental Evaluation

The aim of experimental evaluation was to see how native speakers of English comprehend texts with referential choices, modified in accordance with the algorithm's predictions. If divergent predictions are appropriate referential options, we expect no significant difference in the participants' ability to understand the original and the modified texts, and to answer questions about the referents. If the predicted referential option is inappropriate, we expect that comprehending a modified text is harder. We measure the ease or difficulty of comprehension by the participants' correctness in answering questions about referents, as well as by the participants rating the difficulty of each question.

Materials and Methods

Due to the nature of the natural texts in the corpus, we had to apply a number of restrictions on the material to make it suitable for experimental evaluation. We have selected modified texts from the corpus according to the following criteria:

1. length no less than 140 words, to avoid particularly short texts
2. length not exceeding 260 words, in order to control for the duration of the experiment
3. divergence-containing referential chains that involve at least three anaphors, in order to check the implementation of the flexible script

4. only one divergence per referential chain
5. predicted pronoun in place of an original full NP.

The two latter criteria call for explanatory comments. The decision to select referential chains with one divergence from the original was made in order to have modified and original texts differ in exactly one point, and thus to control for the number of factors involved. The application of the flexible script ensured that, in a given referential chain, after the predicted pronoun all subsequent referential choices did not diverge from the original. Note that the use of the flexible script was still useful: in the earlier comprehension experiment (Khudyakova, 2012) the difficulty of comprehending certain experimental texts could be attributed to the mismatch between the predicted divergent pronoun and the subsequent context. Using the flexible script helped to avoid such situations.

As for the last criterion, we had two reasons for only including the instances of underspecification by the algorithm. First, in the instances of overspecification the exact form of a referential expression (e.g., choice of a nominal lexeme, attributes, etc.) is not generated, and therefore a modified text would contain a referential choice supplied by a human experimenter. Second, this kind of divergence is much more informative: as was discussed in Section "Results", class imbalance leads to the algorithms' general predisposition to predict full NPs.

The resulting experimental set, containing all the texts matching the selection criteria, consisted of six texts. (Note that all of the obvious errors identifed at the stage of expert evaluation were filtered out due to the selection criteria.) We created a modified version of each text: the original full NP was replaced by a predicted pronoun. Then two experimental lists were created, each containing six texts, of which three were in a modified version and three texts were the original ones from the corpus.

Three questions for each text were formulated: one experimental and two control ones. Each experimental question concerned a relevant referential device, that is, one of those for which a pronoun was predicted by the algorithm. WH-words (*who, whom, whose*, or *what*) were used in the experimental questions. One of the control questions was also a WH-question, while the other one was a polar (yes—no) question.

An example of a text can be seen in (5), with the original full NP underlined, followed by the predicted pronoun in brackets. The three questions are provided below with correct responses in parentheses, and the experimental question is underlined.

(5) Milton Petrie, chairman of Petrie Stores Corp. said he has agreed to sell his 15.2% stake in Deb Shops Corp. to Petrie Stores. In a Securities and Exchange Commission filing, Mr. Petrie said that on Oct. 26 Petrie Stores agreed to purchase Mr. Petrie's [his] 2,331,100 Deb Shops shares. The transaction will take place tomorrow. The filing said Petrie Stores of Secaucus, N.J. is purchasing Mr. Petrie's Deb Shops stake as an investment. Although Petrie Stores has considered seeking to acquire the remaining equity of Deb Stores, it has no current intention to pursue such a possibility, the filing said. Philadelphia based Deb Shops said it saw little significance in Mr. Petrie selling his stock to Petrie Stores. We do not look at it and say, 'Oh my God, something is going to happen,' said Stanley Uhr, vice president and corporate counsel. Mr. Uhr said that Mr. Petrie or his company have been accumulating Deb Shops stock for several years, each time issuing a similar regulatory statement. He said no discussions currently are taking place between the two companies.

Whose shares will Petrie stores purchase? (Mr. Petrie's)
Where are Deb Shops based? (Philadelphia)
Does Stanley Uhr work for Petrie stores? (no)

The experiment was run online using the Ibex Farm platform[4]. Each text appeared on the screen one line at a time. In the experiment we presented the texts as closely to their original appearance in the newspaper as possible, so the line length was approximately 40 characters, which matches the size of a column in Wall Street Journal. In order to see the following line of the text a participant had to press a button. Prior text did not disappear from the screen. The self-paced reading design was chosen to ensure that the participants would pay attention to all elements of the experimental texts. After the participants finished reading the text, three questions, one experimental and two control ones, appeared on the screen in a randomized order, one at a time, with the text remaining visible. Since the experimental texts are quite hard for readers (all the texts are rated as "difficult to read" or "college-level" by standard readability metrics, see **Table 9** for details), answering questions without the texts remaining available could result in an excessive rate of errors.

Participants were also asked to rate the difficulty of each question on a 5-point scale, ranging from 1 "very easy" to 5 "very hard".

Twenty four people, including 17 females and 7 males, aged 25 to 36, took part in the experiment. All participants were native speakers of English with college-level education and explicitly stated their willingness to voluntarily participate in the experiment.

Results

Experiment participants answered 18 questions each, that is three questions per text. All participants provided 15 or more correct responses; the number of incorrect responses by participant is summarized in **Table 10**.

Questions can be divided into three groups: experimental questions to original referential expressions, experimental questions to modified (predicted) referential expressions, and control questions. All questions were answered correctly by at least 75% of the participants. The numbers and percentages of correct responses are shown in **Table 11**. The ratings are shown in the right hand part of **Table 11**.

In order to test the equivalence of correct response rates for the three groups of questions we performed the TOST (two one-sided tests) equivalence test (Schuirman, 1987) that treats the difference between groups as a null hypothesis. For the equivalence threshold set at 10%, the test yielded that the experimental groups of responses (modified vs. original referential forms) were

[4] Drummond, A. Ibex Farm. Available at: http://spellout.net/ibexfarm

TABLE 9 | Readability indices for the texts used in the experimental evaluation of referential choice.

Text	Flesch Reading Ease score (Kincaid et al., 1975)	Gunning Fog (Gunning, 1968)	Flesch-Kincaid Grade Level (Kincaid et al., 1975)	The Coleman-Liau Index (Coleman and Liau, 1975)	The SMOG Index (McLaughlin, 1969)	Automated Readability Index (Senter and Smith, 1967)
	30−49: Difficult 50−59: Fairly difficult	Grade level (1 to 12 correspond to school grades, 13 and higher to college levels)				
1	36.0	17.5	14.1	14.0	13.7	15.7
2	58.1	12.3	9.7	10.0	9.4	9.5
3	43.5	17.2	14.1	9.0	12.9	14.1
4	38.0	18.1	15.2	11.0	13.7	15.8
5	36.9	15.6	14.0	11.0	12.7	13.7
6	46.7	13.7	11.7	10.0	12.4	11.1
Average	43.2	15.7	13.1	10.8	12.5	13.3

TABLE 10 | Numbers of correct and incorrect responses given by participants.

Number of incorrect responses (out of 18)	Number of participants
0	6
1	6
2	9
3	3

equivalent ($p = 0.001$, CI 90% [−4.5, 4.5]). This demonstrates that, statistically, the overall perceived correctness does not differ for the original and modified texts. The same test was applied to check for statistical equivalence of correct response rates to experimental questions (about the original expressions), as opposed to responses to control questions. The two groups were proved to be statistically equivalent for the threshold of 10% ($p = 0.001$, CI 90% [−5.1, 3.7]).

We thus did not detect differences between the human understanding of original and predicted referential expressions, and it appears that in the analyzed texts instances of divergent referential choice occur in the situations in which either a full NP or a pronoun is appropriate from a human language user's perspective.

Discussion

The results of both evaluation studies support the idea that the divergent referential options predicted by the algorithm mostly occur in the situations in which a human language user accepts either referential form, or processes both the original and the predicted forms equally well.

Expert evaluation suggests that the majority of discrepancies between the original texts and the algorithm predictions do not result from outright algorithm errors, but rather can be interpreted as equally appropriate referential expressions. The results of the experimental evaluation suggest that, in the selected texts, replacement of a full NP by a pronoun, as predicted by the algorithm, does not lead to increased comprehension difficulty, measured both objectively (correctness of responses) and subjectively (question difficulty ratings). Though the nature of experimental evaluation does not allow us to test all the instances of divergent predictions, the observed results demonstrate that both the original and the predicted referential forms may quite often be equally appropriate.

In experimental evaluation, participants answered questions about the original and modified texts and thus played the role of discourse interpreters, rather than producers. A certain caution must be exercised when extending the experiment results to referential choice, which is a part of discourse production. One might possibly argue that, even if readers allow for more than one referential option, human writers would still perform referential choice in a categorical and deterministic way. Clearly, further experimentation is required, putting human participants in a position closer to that of a discourse producer. Note, however, that the earlier editing experiment reported in Section "Related Work" (Khudyakova et al., 2014) also indicated a strong non-categorical effect in a situation imitating human discourse production.

Overall, we propose that human evaluation of machine learning results provides more precise information about the appropriateness of referential choice prediction than the traditional accuracy measurement. Only human language users can detect whether the divergent referential choices offered by the machine are actually appropriate, and thus provide us with a clear understanding of the algorithm's error rate.

GENERAL DISCUSSION

The approach we used in this study is characterized by several major conceptual elements. First, we mostly focused on the basic referential choice between full and reduced referential devices, also looking occasionally into the second order distinction between two kinds of full NPs: proper names and descriptions. Second, as is suggested by extensive prior research, we took into account a multiplicity of factors affecting referential choice. The factors we have analyzed fall into two major groups: stable referent properties and flexible factors associated with the discourse context, that latter involving several distances from an anaphor to the antecedent. Third, we used a corpus

TABLE 11 | Numbers of correct responses to each question in the experiment and difficulty ratings.

Question group	Question number	Correct responses		Ratings		
		N out of 12	**% of all responses**	**Mean**	**Median**	**Mode**
Experimental questions, original referential expression	1	11	91.67	2.83	3	3
	2	10	83.33	2.67	2.5	2
	3	11	91.67	2.83	3	4
	4	10	83.33	2.75	3	3
	5	11	91.67	2.75	3	3
	6	11	91.67	2.50	2.5	4
Experimental questions, modified referential expression	1	10	83.33	2.50	2.5	3
	2	11	91.67	2.58	3	3
	3	10	83.33	2.75	3	2
	4	11	91.67	2.92	3	3
	5	11	91.67	2.83	3	4
	6	11	91.67	2.58	3	3
		N out of 24	**% of all responses**	**Mean**	**Median**	**Mode**
Control questions	1 yes/no	22	91.67	2.63	2.5	2
	1 WH	23	95.83	2.67	2.5	2
	2 yes/no	22	91.67	2.83	3	3
	2 WH	23	95.83	2.92	3	3
	3 yes/no	20	83.33	2.63	3	3
	3 WH	21	87.50	2.63	3	3
	4 yes/no	21	87.50	2.67	2.5	2
	4 WH	23	95.83	2.58	3	3
	5 yes/no	18	75.00	2.67	3	3
	5 WH	22	91.67	2.67	2.5	2
	6 yes/no	21	87.50	2.67	3	3
	6 WH	22	91.67	2.67	3	3

of texts, sufficient from a statistical point of view. The corpus was annotated for reference and for multiple parameters that potentially can serve as factors of referential choice. Fourth, we employed a cross-methodological approach, combining the corpus-based computational modeling and experimentation with human participants.

Two main findings result from this study, the first one concerned with computational prediction of referential choice, and the second one with the limits of such prediction. Below we summarize them in turn.

Machine learning techniques were used to predict referential choice at each point where an anaphor occured in the corpus texts. In most previous machine learning-based studies of referential choice authors primarily used decision trees. In contrast, our study is characterized by the use of a wide variety of machine learning algorithms, including classifier compositions. Trained models provided almost 90% accurate prediction of referential choices and demonstrated that machine learning algorithms can imitate referential choices made by human language users with substantial success. We also explored the cumulative and individual contribution of various factors to the resulting referential choice.

In spite of the relatively successful modeling results, prediction accuracy did not approach 100%, and this raised the question of whether complete accuracy is attainable. In order to address this question, we used experimentation with human participants. We submitted the results of modeling to human judgment and assessed the divergences between the original and predicted referential choices as appropriate or inappropriate from the language users' point of view. Experiment results suggest that there are numerous instances in which referential options are equally appropriate for human participants. Accordingly, many of the algorithm's failures to predict referential choice exactly as in original texts may be due not to plain error but to inherently not fully categorical nature of referential choice. Even a perfect algorithm (or, for that matter, another human language user, or even the same language user on a different occasion) could not be expected to necessarily make the choice once implemented in a text. In other words, a certain degree of variation must be built into a realistic model of referential choice. Even if the algorithm learns to imitate non-categorical referential choice (cf. examples of non-deterministic REG algorithms in van Deemter et al., 2012), mismatches between the algorithm's prediction and the original text would be inevitable.

A few notes are in order regarding the theoretical context of this study. Following many other students of discourse reference (Chafe, 1976; Givón, 1983, and numerous later studies), we suppose that referential choice is immediately governed by a referent's status in the speaker's cognition. In particular, more attenuated forms of reference are used when the referent is more salient or more activated for the speaker/writer. According to the model assumed in this study, the cognitive component responsible for referential choice is activation in working memory, and different levels of referent activation are responsible for using either a reduced or a full referential device (**Figure 1**). In this model, the linguistic factors affecting referential choice are interpreted as activation factors. Operating in conjunction, they contribute to a referent's current activation, which, in turn, determines referential choice. In some of our previous studies (Kibrik, 1996, 1999) referent's summary activation was computed numerically and served as an explanatory component. In the present study, the activation component is not technically implemented, as standard computer modeling techniques only provide information on the mappings from linguistic factors to referential choice. Nevertheless, we believe that it is important to keep the cognitively realistic picture in mind, even if one has to remain at the level of form-to-form mappings.

The same applies to the issue of incomplete categoricity of referential choice. We demonstrated that human language users accept more than one referential option in many contexts. One can remain at the level of such observation, but it is interesting to inquire into the causes of non-categoricity. The cognitive model assumed in this study offers a plausible explanation to this phenomenon: variation of referential options occurs in the case of intermediate referent activation; see an amendment to our cognitive model in **Figure 4**. The conclusion on the not fully categorical nature of referential choice appears particularly relevant in the contemporary context of reference studies. There is a growing interest to the variation in the use of referential expressions both in computational studies and in experimental psycholinguistics (see multiple references in Sections "Discussion: Referential Choice Is Not Always Categorical" and "Related Work"), and this study contributes to the duscussion of the possible kinds and causes of such variation. The outcome of this study thus provides support to the previously expressed idea that "non-determinism should be an important property of a psychologically realistic algorithm" (van Deemter et al., 2012, p. 19).

There are several avenues for further development of the present approach in future research. As pointed out above, machine learning algorithms normally only give access to the input layer (activation factors) and the output layer (referential choice prediction), the internal working of the algorithms remaining hidden. We would like to reinstate the cognitive interpretation that is the degrees of activation that result from the activation factors in conjunction and directly map onto referential choice. One way how this can be done is associated with some algorithms' (e.g., logistic regression) capacity to evaluate the contribution of various factors and the certainty of prediction, which can be interpreted as activation factors and summary activation level, respectively. This can also be a path to training the algorithms to model non-categorical referential choice.

The cognitive model shown in **Figure 1** is simplified in that it leaves out the filter of referential conflict, or ambiguity, that modulates referential choice after referent activation is computed (see Fedorova et al., 2010a; Kibrik, 2011; Fedorova, 2014). Sometimes a reduced referential device is filtered out because it creates a potential ambiguity for the addressee, for the reason that there is more than one highly activated referent. As of now, some of the referential conflict-related factors, such as gender and distance in all markables, are taken into account in our modeling study, but they are interspersed among the activation factors. We intend to clarify the distinction between referent activation and the referential conflict filter in future research.

In our modeling study, there is probably space for tuning up certain activation factors, which may lead to some further improvement of prediction. As was pointed out in Section "Human Evaluation", we detected some algorithm errors, such as overspecification in the context of syntactic anaphora or underspecification at the beginning of a referential chain. These kinds of errors can be fixed by modifying the set of factors.

The set of factors responsible for the basic referential choice turned out quite efficient in predicting the second-order choice between descriptions and proper names (end of Section "Results"). A more focused search for factors directly related to this choice is in order. Also, the proposed approach can be extended to further details of referential choice, such as varieties of attributes in descriptions, as well as less frequent referential options, e.g., demonstratives. We also believe that our approach can be used in the domains of language production other than referential choice.

In this study we looked at written discourse, as a well-controlled testing ground for sharpening the methods of cognitive and computational modeling and as the material easily lending itself to various kinds of manipulation. We assume that, in spite of the special character of newspaper texts, written discourse is created on the basis of general cognitive principles of discourse production, including referential choice, and that the discovered regularities can in principle be extended to other types of language use. Nowadays, linguistic research is opening up new horizons, including interest in interactive face-to-face communication, visual context, and multimodality. All of these developments are also relevant to the study of referential choice, see e.g., Janarthanam and Lemon (2010), Viethen (2011), de Ruiter et al. (2012), and Hoetjes et al. (2015). The theoretical and methodological approach, developed here on the basis of written texts, can also be applied to a wide range of discourse types, including various genres, spoken discourse, conversation, and multimodal interaction.

AUTHOR CONTRIBUTIONS

AK has conceived the general design of the study, developed the theoretical framework, selected the corpus for analysis, put together the team, allocated the assignments to coauthors, formulated the general structure of the paper, wrote Sections "Introduction" and "General Discussion", drafted "Corpus-Based Modeling", and edited the whole text. MK worked substantially on the corpus, developed the annotation scheme, organized the work of student assistants, designed the experimental part, conducted the experiment, and wrote Section “Experimental Studies of Referential Variation”. GB provided expertise on machine learning, conducted multiple modeling studies, helped to plan the whole study, wrote parts of Sections “Corpus-Based Modeling” and “Experimental Studies of Referential Variation”. AL provided expertise on discourse annotation, natural language generation, psycholinguistic experimentation, wrote literature surveys, complied the bibliography, performed technical editing of the paper, provided big input on all aspects of the paper. DZ worked on the corpus, organized the work of student assistants, wrote “Corpus-Based Modeling”, performed technical editing of the paper, provided big input on all aspects of the paper. All coauthors participated in developing the design of the study, acquiring the data, analyzing data, writing up the manuscript, contributed to multiple manuscript revision throughout all of stages of paper preparation.

FUNDING

The research underlying this paper was supported by the Russian Foundation for Basic Research (grant #14-06-00211). The work of MK was also supported by the Basic Research Program of the National Research University Higher School of Economics (HSE) and within the framework of a subsidy by the Russian Academic Excellence Project “5-100”.

ACKNOWLEDGMENTS

The authors would like to thank Olga Fedorova and Natalia Loukachevitch for their highly valuable assistance in the preparation of this paper. We are also grateful to the journal reviewers and Kees van Deemter for their useful comments. We would like to acknowledge the support of the German Research Foundation (Deutsche Forschungsgemeinschaft) and Open Access Publishing Fund of University of Potsdam.

REFERENCES

Anderson, J., and Hastie, R. (1974). Individuation and reference in memory: proper names and definite descriptions. *Cogn. Psychol.* 6, 495–514. doi: 10.1016/0010-0285(74)90023-1

Ariel, M. (1990). *Accessing NP Antecedents*. London: Routledge.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Process.* 31, 137–162. doi: 10.1207/S15326950DP3102_02

Arnold, J. E. (2015). Women and men have different discourse biases for pronoun interpretation. *Discourse Process.* 52, 77–110. doi: 10.1080/0163853X.2014.946847

Arnold, J. E., and Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: everyone counts. *J. Mem. Lang.* 56, 521–536. doi: 10.1016/j.jml.2006.09.007

Arnold, J. E., Nozari, N., and Thompson-Schill, S. L. (2014). Stimulation of left prefrontal cortex increases discourse connectedness and reduced references. *Paper presented at RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production,* Edinburgh.

Arutjunova, N. D. (1977). “Nominacija i tekst [Nomination and text],” in *Jazykovaja nominacija (tipy naimenovanij)*, eds B. A. Serebrennikov and A. Ufimceva (Moscow: Nauka), 304–357.

Automatic Content Extraction (2004). *Annotation Guidelines for Entity Detection and Tracking (EDT) Version 4.2.6*. Available at: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-edt-v4.2.6.pdf

Awh, E., and Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends Cogn. Sci.* 5, 119–126. doi: 10.1016/S1364-6613(00)01593-X

Awh, E., Vogel, E. K., and Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience* 139, 201–208. doi: 10.1016/j.neuroscience.2005.08.023

Baumann, P., Clark, B., and Kaufmann, S. (2014). “Overspecification and the cost of pragmatic reasoning about referring expressions,” in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, eds P. Bello, M. Guarini, M. McShane, and B. Scassellati (Austin, TX: Cognitive Science Society), 1898–1903.

Belz, A., and Kow, E. (2010). “The GREC challenges 2010: overview and evaluation results,” in *Proceedings of the 6th International Natural Language Generation Conference*, eds J. Kelleher, B. Mac Namee, and I. van der Sluis (Trim: Association for Computational Linguistics), 219–229.

Belz, A., Kow, E., Viethen, J., and Gatt, A. (2008). “The GREC challenge: overview and evaluation results,” in *Proceedings of the Fifth International Natural Language Generation Conference, Salt Fork, Ohio*, eds M. White, C. Nakatsu, and D. McDonald (Stroudsburg, PA: Association for Computational Linguistics), 183–191.

Belz, A., Kow, E., Viethen, J., and Gatt, A. (2009). “The GREC main subject reference generation challenge 2009: overview and evaluation results,” in *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, eds A. Belz, R. Evans, and S. Varges (Morristown, NJ: Association for Computational Linguistics), 79–87.

Belz, A., Kow, E., Viethen, J., and Gatt, A. (2010). “Generating referring expressions in context: the GREC task evaluation challenges,” in *Empirical Methods in Natural Language Generation*, eds E. Krahmer and M. Theune (Berlin: Springer), 294–328.

Belz, A., and Varges, S. (2007). “Generation of repeated references to discourse entities,” in *Proceedings of the Eleventh European Workshop on Natural Language Generation*, (Stroudsburg, PA: Association for Computational Linguistics), 9–16.

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1023/A:1018046112532

Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). “A centering approach to pronouns,” in *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, (Stroudsburg, PA: Association for Computational Linguistics), 155–162.

Carlson, L., Marcu, D., and Okurowski, M. (2002). *RST Discourse Treebank*. Philadelphia, PA: Linguistic Data Consortium.

Chafe, W. L. (1976). “Givenness, contrastiveness, definiteness, subjects, topics and point of view,” in *Subject and Topic*, ed. C. N. Li (New York, NY: Academic Press), 27–55.

Chafe, W. L. (1994). *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago, IL: University of Chicago Press.

Cheng, H., Poesio, M., Henschel, R., and Mellish, C. (2001). “Corpus-based NP modifier generation,” in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, Stroudsburg, PA, 1–8.

Chinchor, N., and Robinson, P. (1997). “MUC-7 named entity task definition,” in *Proceedings of the 7th Conference on Message Understanding*, Fairfax, VA, 29.

Chinchor, N., and Sundheim, B. (1995). “Message Understanding Conference (MUC) tests of discourse processing,” in *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford, CA, 21–26.

Clancy, P. M. (1980). “Referential choice in English and Japanese narrative discourse,” in *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, ed. W. L. Chafe (Norwood, NJ: Ablex), 127–201.

Coleman, M., and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *J. Appl. Psychol.* 60, 283–284. doi: 10.1037/h0076540

Cornish, F. (1999). *Anaphora, Discourse, and Understanding: Evidence from English and French*. Oxford: Oxford University Press.

Cowan, N. (1995). *Attention and Memory: An Integrated Framework*. Oxford: Oxford University Press.

Dahl, Ö., and Fraurud, K. (1996). "Animacy in grammar and discourse," in *Reference and Referent Accessibility*, eds T. Fretheim and J. K. Gundel (Amsterdam: John Benjamins), 47–64.

Dale, R. (1992). *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. Cambridge: MIT Press.

Dale, R., and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cogn. Sci.* 19, 233–263. doi: 10.1207/s15516709cog1902_3

Dale, R., and Viethen, J. (2010). "Attribute-centric referring expression generation," in *Empirical Methods in Natural Language Generation*, eds E. Krahmer and M. Theune (Berlin: Springer), 163–179.

de Ruiter, J. P., Bangerter, A., and Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: investigating the tradeoff hypothesis. *Top. Cogn. Sci.* 4, 232–248. doi: 10.1111/j.1756-8765.2012.01183.x

Dethlefs, N. (2014). Context-sensitive natural language generation: from knowledge-driven to data-driven techniques. *Lang. Linguist. Compass* 8, 99–115. doi: 10.1111/lnc3.12067

Dethlefs, N., and Cuayáhuitl, H. (2011). "Hierarchical reinforcement learning and hidden Markov models for task-oriented natural language generation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers* - Vol. 2, (Stroudsburg, PA: Association for Computational Linguistics), 654–659.

Di Fabbrizio, G., Stent, A. J., and Bangalore, S. (2008). "Trainable speaker-based referring expression generation," in *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, Manchester, eds A. Clark and K. Toutanova (Stroudsburg, PA: Association for Computational Linguistics), 151–158.

Efimova, Z. V. (2006). *Referencial'naja Struktura Narrativa v Japonskom Jazyke (v Sopostavlenii s Russkim) [Referential Structure of Narrative in Japanese, as Compared to Russian]*. Ph.D. thesis, Institute of Linguistics, Russian State University for the Humanities.

Enfield, N. J., and Stivers, T. (2007). *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives*. Cambridge: Cambridge University Press.

Engle, R. W., and Kane, M. J. (2004). "Exectutive attention, working memory capactiy, and a two-factor theory of cognitive control," in *The Psychology of Learning and Motivation*, Vol. 44, ed. B. Ross (New York, NY: Elsevier), 145–199.

Engonopoulos, N., and Koller, A. (2014). "Generating effective referring expressions using charts," in *Proceedings of the INLG and SIGDIAL 2014 Joint Session*, (Stroudsburg, PA: Association for Computational Linguistics).

Fedorova, O. V. (2014). The role of potential referential conflict in the choice of a referring expression. *Russ. J. Cogn. Sci.* 1, 6–21.

Fedorova, O. V., Delikishkina, E. A., Malyutina, S. A., Uspenskaya, A. M., and Feyn, A. A. (2010a). "Eksperimental'nyj podxod k issledovaniju referencii v diskurse: interpretatcija anaforicheskogo mestoimenija v zavisimosti ot ritoricheskogo rasstoianija do ego antetcedenta [Experimental approach to reference in discourse: interpretation of anaphoric pronoun depending on the rhetorical distance to its antecedent]," in *Proceedings of the Papers from the Annual International Conference "Dialogue," Bekasovo: Computational Linguistics and Intellectual Technologies*, eds A. E. Kibrik, V. I. Belikov, B. V. Dobrov, D. O. Dobrovolsky, L. M. Zakharov, I. M. Zatsman, et al. (Moscow: Izdatel'stvo RGGU), 525–531.

Fedorova, O. V., Delikishkina, E. A., and Uspenskaya, A. M. (2010b). "Experimental approach to reference in discourse: working memory capacity and language comprehension in Russian," in *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, (Sendai: Tohoku University), 125–132.

Fedorova, O. V., Delikishkina, E. A., and Uspenskaya, A. M. (2012). "Empiricheskie issledovanija referencii v diskurse: rol' ritoricheskoj struktury v processax porozhdenija i ponimanija referentcial'nogo vyrazhenija [Empirical studies of reference in discourse: the role of rhetorical structure in the processess of generation and understanding referring expressions]," in *Kognitivnye issledovanija [Cognitive Studies]*, eds A. A. Kibrik and T. V. Chernigovskaya (Moscow: Institut Psixologii), 230–242.

Ferreira, T. C., Krahmer, E., and Wubben, S. (2016). "Towards more variation in text generation: developing and evaluating variation models for choice of referential form," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin.

Fox, B. A. (1987). Anaphora in popular written English narratives. *Coherence Ground. Discourse* 11, 157–174. doi: 10.1075/tsl.11.09fox

Freund, Y., and Schapire, R. E. (1996). "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning*, Bari. ed. L. Saitta (San Mateo, CA: Morgan Kaufmann), 148–156.

Fukumura, K., Hyönä, J., and Scholfield, M. (2013). Gender affects semantic competition: the effect of gender in a non-gender-marking language. *J. Exp. Psychol. Learn. Mem. Cogn.* 39, 1012–1021. doi: 10.1037/a0031215

Fukumura, K., and van Gompel, R. P. G. (2010). Choosing anaphoric expressions: do people take into account likelihood of reference? *J. Mem. Lang.* 62, 52–66. doi: 10.1016/j.jml.2009.09.001

Fukumura, K., and van Gompel, R. P. G. (2011). The effect of animacy on the choice of referring expression. *Lang. Cogn. Process.* 26, 1472–1504. doi: 10.1080/01690965.2010.506444

Fukumura, K., and van Gompel, R. P. G. (2015). Effects of order of mention and grammatical role on anaphor resolution. *J. Exp. Psychol. Learn. Mem. Cogn.* 41, 501–525. doi: 10.1037/xlm0000041

Garrod, S. C. (2011). "Referential processing in monologue and dialogue with and without access to real world referents," in *The Processing and Acquisition of Reference*, eds E. Gibson and N. J. Pearlmutter (Cambridge, MA: MIT Press), 273–294.

Gatt, A., and Belz, A. (2008). "Attribute selection for referring expression generation: new algorithms and evaluation methods," in *Proceedings of the Fifth International Natural Language Generation Conference,* Salt Fork, OH. eds M. White, C. Nakatsu, and D. McDonald (Stroudsburg, PA: Association for Computational Linguistics), 50–58.

Gatt, A., Krahmer, E., van Deemter, K., and van Gompel, R. P. G. (2014). Models and empirical data for the production of referring expressions. *Lang. Cogn. Neurosci.* 29, 899–911. doi: 10.1080/23273798.2014.933242

Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hove: Psychology Press.

Givón, T. (1983). "Topic continuity in discourse: an introduction," in *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, ed. T. Givón (Amsterdam: John Benjamins), 3–41.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cogn. Sci.* 17, 311–347. doi: 10.1207/s15516709cog1703_1

Greenbacker, C. F., and McCoy, K. F. (2009). Feature selection for reference generation as informed by psycholinguistic research. *Paper Presented at the Production of Referring Expressions Conference: Bridging the Gap between Computational and Empirical Approaches to Reference (PRE-CogSci)*, Amsterdam.

Grimes, J. E. (ed.) (1978). *Papers on Discourse*. Dallas, TX: Summer Institute of Linguistics.

Grishman, R., and Sundheim, B. (1995). "Design of the MUC-6 evaluation," in *Proceedings of the 6th Conference on Message Understanding,* Columbia, MD, 1–11. doi: 10.3115/1072399.1072401

Grüning, A., and Kibrik, A. A. (2005). "Modeling referential choice in discourse: a cognitive calculative approach and a neural network approach," in *Anaphora Processing: Linguistic, Cognitive and Computational Modeling*, eds A. H. Branco, T. McEnery, and R. Mitkov (Amsterdam: John Benjamins), 163–198.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307. doi: 10.2307/416535

Gundel, J. K., Hedberg, N., and Zacharski, R. (2012). Underspecification of cognitive status in reference production: some empirical predictions. *Top. Cogn. Sci.* 4, 249–268. doi: 10.1111/j.1756-8765.2012.01184.x

Gunning, R. (1968). *The Technique of Clear Writing*. New York, NY: McGraw-Hill.

Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., et al. (2009). The WEKA data mining software: an update. *SIGKDD Explor.* 11, 10–18. doi: 10.1145/1656274.1656278

Heller, D., Gorman, K. S., and Tanenhaus, M. K. (2012). To name or to describe: shared knowledge affects referential form. *Top. Cogn. Sci.* 4, 290–305. doi: 10.1111/j.1756-8765.2012.01182.x

Helmbrecht, J. (2009). On the typology of proper names. *Paper Presented at the 8th Conference of the Association for Linguistic Typology*, Berkeley, CA.

Hendriks, P. (2014). "The speaker's perspective," in *Asymmetries between Language Production and Comprehension*, ed. P. Hendriks (Netherlands: Springer), 123–152.

Hendriks, P. (2016). Cognitive modeling of individual variation in reference production and comprehension. *Front. Psychol.* 7:506. doi: 10.3389/fpsyg.2016.00506

Hendriks, P., Koster, C., and Hoeks, J. C. J. (2014). Referential choice across the lifespan: why children and elderly adults produce ambiguous pronouns. *Lang. Cogn. Neurosci.* 29, 391–407. doi: 10.1080/01690965.2013.766356

Hinds, J. (ed.). (1978). *Anaphora in Discourse*. Edmonton, AB: Linguistic Research.

Hirschberg, J. (1993). Pitch accent in context predicting intonational prominence from text. *Artif. Intell.* 63, 305–340. doi: 10.1016/0004-3702(93)90020-C

Hobbs, J. R. (1985). *On the Coherence and Structure of Discourse*. Report No. CSLI-85-37. Stanford, CA: Stanford University, Center for the Study of Language and Information.

Hoetjes, M., Krahmer, E., and Swerts, M. (2015). On what happens in gesture when communication is unsuccessful. *Speech Commun.* 72, 160–175. doi: 10.1016/j.specom.2015.06.004

Hovy, E. H. (1993). Automated discourse generation using discourse structure relations. *Artif. intell.* 63, 341–385. doi: 10.1016/0004-3702(93)90021-3

Hughes, E. M., and Allen, S. E. M. (2013). The effect of individual discourse-pragmatic features on referential choice in child English. *J. Pragmat.* 56, 15–30. doi: 10.1016/j.pragma.2013.05.005

Janarthanam, S., and Lemon, O. (2010). "Adaptive referring expression generation in spoken dialogue systems: evaluation with real users," in *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (Stroudsburg, PA: Association for Computational Linguistics), 124–131.

Jordan, P. W., and Walker, M. A. (2005). Learning content selection rules for generating object descriptions in dialogue. *J. Artif. Intell. Res.* 24, 157–194.

Joshi, A. K., Prasad, R., and Miltsakaki, E. (2006). "Anaphora resolution: centering theory approach," in *Encyclopedia of Language & Linguistics*, Vol. 1, eds E. K. Brown, R. E. Asher, and J. M. Y. Simpson (Amsterdam: Elsevier), 223–230.

Kaiser, E. (2008). Multiple dimensions in anaphor resolution. *Paper Presented at the Workshop on General Cognition and Language Processing, 3rd International Conference on Cognitive Science*, Moscow.

Kaiser, E., Li, D. C. H., and Holsinger, E. (2011). "Exploring the lexical and acoustic consequences of referential predictability," in *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011): Anaphora Processing and Applications*, Faro, eds I. Hendrickx, S. L. Devi, A. Branco, and R. Mitkov (Springer: Berlin Heidelberg), 171–183.

Kameyama, M. (1999). "Stressed and unstressed pronouns: complementary preferences," in *Focus: Linguistic, Cognitive and Computational Perspectives*, eds P. Bosch and R. van der Sandt (Cambridge: Cambridge University Press), 306–321.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. Stanford, CA: CSLI Publications.

Kelly, C., Copestake, A., and Karamanis, N. (2009). "Investigating content selection for language generation using machine learning," in *Proceedings of the 12th European Workshop on Natural Language Generation,* Athens (Stroudsburg, PA: Association for Computational Linguistics), 130–137.

Khan, I. H., van Deemter, K., and Ritchie, G. (2012). Managing ambiguity in reference generation: the role of surface structure. *Top. Cogn. Sci.* 4, 211–231. doi: 10.1111/j.1756-8765.2011.01167.x

Khudyakova, M. V. (2012). "Akkuratnost' modelirovanija referencial'nogo vybora: ocenka chitateljami [Accuracy of referential choice modeling: reader evaluation]," in *Proceedings of the Fifth International Conference on Cognitive Science*, Vol. 2. *Abstracts,* eds Yu. I. Aleksandrov, K. V. Anokhin, B. M. Velichkovsky, A. V. Dubasova, A. A. Kibrik, A. K. Krylov, et al. (Kaliningrad: Russian Association for Cognitive Research), 688–690.

Khudyakova, M. V., Kibrik, A. A., and Dobrov, G. B. (2014). "Nekategoricheskij referencial'nyj vybor [Non-categorical referential choice]," in *Proceedings of the Sixth International Conference on Cognitive Science*, Kaliningrad.

Kibrik, A. A. (1996). "Anaphora in Russian narrative discourse: a cognitive calculative account," in *Studies in anaphora*, ed. B. Fox (Amsterdam: John Benjamins), 255–304.

Kibrik, A. A. (1999). "Cognitive inferences from discourse observations: reference and working memory," in *Proceedings of the 5th International cognitive linguistics conference: Discourse Studies in Cognitive Linguistics*, eds K. van Hoek, A. A. Kibrik, and L. Noordman (Amsterdam: John Benjamins), 29–52.

Kibrik, A. A. (2011). *Reference in Discourse*. Oxford: Oxford University Press.

Kibrik, A. A., Dobrov, G. B., Zalmanov, D. A., Linnik, A. S., and Loukachevitch, N. V. (2010). "Referencial'nyj vybor kak mnogofaktornyj verojatnostnyj process [Referential choice as a multi-factorial probabilistic process]," in *Proceedings of the Papers from the Annual International Conference "Dialogue" (2010)*, Bekasovo: *Computational Linguistics and Intellectual Technologies*, ed. A. E. Kibrik (Moscow: Izdatel'stvo RGGU), 173–181.

Kibrik, A. A., and Krasavina, O. N. (2005). A corpus study of referential choice: the role of rhetorical structure. *Papers from the Annual International Conference "Dialogue": Computational Linguistics and Intellectual Technologies*, eds I. M. Kobozeva, A. S. Narin'jani, and V. P. Selegey (Moscow: Nauka), 561–569.

Kincaid, J. P., Fishburne, R. P. Jr., Rogers, R. L., and Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* Research Branch Report No. 8-75. Millington, TN: Naval Technical Training.

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *J. Pragmat.* 43, 3231–3250. doi: 10.1016/j.pragma.2011.06.008

Krahmer, E., Theune, M., Viethen, J., and Hendrickx, I. (2008). "Graph: the costs of redundancy in referring expressions," in *Proceedings of the Fifth International Natural Language Generation Conference,* Salt Fork, OH, eds M. White, C. Nakatsu, and D. McDonald (Stroudsburg, PA: Association for Computational Linguistics), 227–229.

Krahmer, E., and van Deemter, K. (2012). Computational generation of referring expressions: a survey. *Comput. Linguist.* 38, 173–218. doi: 10.1162/COLI_a_00088

Krasavina, O. N. (2006). "Multi-factorial choices in speaking," in *The Second Biennial Conference on Cognitive Science*, eds B. M. Velichkovsky, T. V. Chernigovskaya, Yu. I. Aleksandrov, and D. N. Akhapkin (Saint-Petersburg: Saint-Petersburg State University, Philological Faculty), 86–87.

Krasavina, O. N., and Chiarcos, C. (2007). "PoCoS: potsdam coreference scheme," in *Proceedings of the Linguistic Annotation Workshop*, Prague, (Stroudsburg, PA: Association for Computational Linguistics), 156–163.

Linnik, A., and Dobrov, G. (2011). "Protagonism as a factor affecting referential choice in discourse," in *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium*, Faro.

Longadge, R., Dongre, S. S., and Malik, L. (2013). Class imbalance problem in data mining: review. *Int. J. Comput. Sci. Netw.* 2, 83–87.

Loukachevitch, N. V., Dobrov, G. B., Kibrik, A. A., Khudyakova, M. V., and Linnik, A. S. (2011). "Factors of referential choice: computational modeling," in *Proceedings of the Papers from the Annual International Conference "Dialogue" (2011): Computational Linguistics and Intellectual Technologies*, ed. A. E. Kibrik. (Moscow: Izdatel'stvo RGGU), 458–467.

Malouf, R. (2000). "The order of prenominal adjectives in natural language generation," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong*, (Stroudsburg, PA: Association for Computational Linguistics), 85–92. doi: 10.3115/1075218.1075230

Mann, W. C., and Thompson, S. A. (1987). "Rhetorical structure theory: description and construction of text structures," in *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, ed. G. Kempen (Dordrecht: Nijhoff (Kluwer)), 85–95.

Marcu, D. (2003). *Discourse Structures: Trees or Graphs?* Available at: http://www.isi.edu/∼marcu/discourse/Discourse%20structures.htm

Marslen-Wilson, W., Levy, E., and Tyler, L. K. (1982). "Producing interpretable discourse: the establishment and maintenance of reference," in *Speech, Place, and Action. Studies in deixis and related topics*, eds R. J. Jarvella and W. Klein (Chichester: Wiley), 339–378.

McCoy, K. F., and Strube, M. (1999). "Generating anaphoric expressions: pronoun or definite description," in *Proceedings of ACL workshop on Discourse and Reference Structure, University of Maryland,* College Park, MD, 63–71.

McLaughlin, G. H. (1969). SMOG grading: a new readability formula. *J. Read.* 12, 639–646.

Miltsakaki, E., Prasad, R., Joshi, A. K., and Webber, B. L. (2004). "The Penn discourse treebank," in *Proceedings of the 4th International Conference on Language Resources and Evaluation,* eds M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, et al. (Paris: European Language Resources Association), 2237–2240.

Müller, C., and Strube, M. (2006). "Multi-level annotation of linguistic data with MMAX2," in *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*, eds S. Braun, K. Kohn, and J. Mukherjee (Frankfurt: Peter Lang), 197–214.

Nicolae, C., Nicolae, G., and Roberts, K. (2010). "C-3: coherence and coreference corpus," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation, Valletta*, eds N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al. (Paris: European Language Resources Association), 136–143.

Nieuwland, M. S., and van Berkum, J. J. A. (2006). Individual differences and contextual bias in pronoun resolution: evidence from ERPs. *Brain Res.* 1118, 155–167. doi: 10.1016/j.brainres.2006.08.022

Paducheva, E. V. (1965). O strukture abzaca [On the structure of paragraph]. *Trudy po znakovym sistemam* 2, 284–292.

Poesio, M. (2000). "Annotating a corpus to develop and evaluate discourse entity realization algorithms: issues and preliminary results," in *Proceedings of the Second International Conference on Language Resources and Evaluation,* Athens.

Poesio, M. (2004). "Discourse annotation and semantic annotation in the GNOME corpus," in *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, Barcelona, eds B. Webber and D. Byron (Stroudsburg, PA: Association for Computational Linguistics), 72–79.

Poesio, M., and Artstein, R. (2008). "Anaphoric annotation in the ARRAU corpus," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation,* Marrakech, eds N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al. (Paris: European Language Resources Association), 1170–1174.

Poesio, M., Henschel, R., Hitzeman, J., and Kibble, R. (1999). "Statistical NP generation: a first report," in *Proceedings of the European Summer School in Logic, Language and Information Workshop on NP Generation*, Utrecht.

Polanyi, L. (1985). "A theory of discourse structure and discourse coherence," in *Proceedings of the Papers from the General Session at the 21st Regional Meeting of the Chicago Linguistics Society*, Chicago, eds P. D. Kroeber, W. H. Eilfort, and K. L. Peterson, (Chicago: Chicago Linguistics Society), 306–322.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., et al. (2008). "The Penn discourse treebank 2.0," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, eds N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, et al. (Paris: European Language Resources Association), 2961–2968.

Reiter, E., and Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge: Cambridge University Press.

Repovš, G., and Bresjanac, M. (2006). Cognitive neuroscience of working memory: a prologue. *Neuroscience* 139, 1–3. doi: 10.1016/j.neuroscience.2005.12.007

Rohde, H., and Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Lang. Cogn. Neurosci.* 29, 912–927. doi: 10.1080/01690965.2013.854918

Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Min. Knowl. Discov.* 1, 317–327. doi: 10.1023/A:1009752403260

Schapire, R. E. (2003). The boosting approach to machine learning: an overview. *Nonlinear Estimation Classif.* 171, 149–171.

Schuirman, D. J. (1987). A comparison of the two one-sided tests procedure and power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* 15, 657–680. doi: 10.1007/BF01068419

Scott, K. (2013). Pragmatically motivated null subjects in English: a relevance theory perspective. *J. Prag.* 53, 68–83. doi: 10.1016/j.pragma.2013.04.001

Seleznev, M. (1987). "Referencija i nominacija [Reference and nomination]," in *Modelirovanie Jazykovoj Dejatel'nosti v Intellektual'nyx Sistemax*, eds A. E. Kibrik and A. S. Narin'jani (Moscow: Nauka), 64–78.

Senter, R. J., and Smith, E. A. (1967). *Automated Readability Index.* Technical Report AMRLTR-66-220. Cincinnati, OH: University of Cincinnati.

Shipstead, Z., Harrison, T. L., and Engle, R. W. (2015). Working memory capacity and the scope and control of attention. *Atten. Percept. Psychophys.* 77, 1863–1880. doi: 10.3758/s13414-015-0899-0

Siddharthan, A., and Copestake, A. (2004). "Generating referring expressions in open domains," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, (Stroudsburg, PA: Association for Computational Linguistics), 407.

Stede, M., and Neumann, A. (2014). "Potsdam commentary corpus 2.0: annotation for discourse research," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation,* Reykjavik, eds N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, et al. (Paris: European Language Resources Association), 925–929.

Stent, A., and Bangalore, S. (eds). (2014). *Natural Language Generation in Interactive Systems*. Cambridge: Cambridge University Press.

Stirling, L. (2001). The multifunctionality of anaphoric expressions: a typological perspective. *Aust. J. Linguist.* 21, 7–23. doi: 10.1080/07268600120042435

Strube, M., and Wolters, M. (2000). "A probabilistic genre-independent model of pronominalization," in *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, ed. J. Wiebe (Stroudsburg, PA: Association for Computational Linguistics), 18–25.

Taboada, M., and Mann, W. C. (2006). Rhetorical structure theory: looking back and moving ahead. *Discourse Stud.* 8, 423–460. doi: 10.1177/1461445606064836

Tetreault, J. R. (2001). A corpus-based evaluation of centering and pronoun resolution. *Comput. Linguist.* 27, 507–520. doi: 10.1162/089120101753342644

Tomlin, R. S. (1987). Linguistic reflections of cognitive events. *Coherence Ground. Discourse* 11, 455–479. doi: 10.1075/tsl.11.20tom

Toole, J. (1996). "The effect of genre on referential choice," in *Reference and Referent Accessibility*, eds T. Fretheim and J. K. Gundel (Amsterdam: John Benjamins), 263–290.

van Deemter, K. (2002). Generating referring expressions: boolean extensions of the incremental algorithm. *Comput. Linguist.* 28, 37–52. doi: 10.1162/089120102317341765

van Deemter, K., Gatt, A., van Gompel, R. P. G., and Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. *Top. Cogn. Sci.* 4, 166–183. doi: 10.1111/j.1756-8765.2012.01187.x

van Rij, J., van Rijn, H., and Hendriks, P. (2011). "WM load influences the interpretation of referring expressions," in *Proceedings of the 2nd workshop on Cognitive Modeling and Computational Linguistics*, eds F. Keller and D. Reiter (Stroudsburg, PA: Association for Computational Linguistics), 67–75.

Vieira, R., and Poesio, M. (1999). "Processing definite descriptions in corpora," in *Corpus-based and Computational Approaches to Discourse Anaphora*, eds S. Botley and T. McEnery (Amsterdam: John Benjamins), 189–212.

Viethen, H. A. E. (2011). *The Generation of Natural Descriptions: Corpus-Based Investigations of Referring Expressions in Visual Domains*. Ph.D. thesis, Macquarie University, Sydney, NSW.

Viethen, J., and Dale, R. (2006a). "Algorithms for generating referring expressions: do they do what people do?," in *Proceedings of the Fourth International Natural Language Generation Conference,* Sydney, NSW, 63–70.

Viethen, J., and Dale, R. (2006b). "Towards the evaluation of referring expression generation," in *Proceedings of the 4th Australasian Language Technology Workshop (ALTW 2006),* Sydney, NSW, 115–122.

Viethen, J., Dale, R., and Guhe, M. (2011). "Serial dependency: is it a characteristic of human referring expression generation?," in *Proceedings of the Workshop on Production of Referring Expressions: Bridging the Gap between Empirical, Computational and Theoretical Approaches to Reference (Pre-CogSci 2011),* Boston, MA.

Vogels, J., Krahmer, E., and Maes, A. (2014). How cognitive load influences speakers' choice of referring expressions. *Cogn. Sci.* 39, 1396–1418. doi: 10.1111/cogs.12205

Walker, M. A., Rainbow, O. C., and Rogati, M. (2002). Training a sentence planner for spoken dialogue using boosting. *Comput. Speech Lang.* 16, 409–433. doi: 10.1016/S0885-2308(02)00027-X

Wolf, F., and Gibson, E. (2003). *A Response to Marcu (2003). Discourse Structures: Trees or Graphs*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?

Wolf, F., Gibson, E., Fisher, A., and Knight, M. (2003). *A Procedure for Collecting a Database of Texts Annotated with Coherence Relations. Database Documentation*. Available at: http://www.ling.ohio-state.edu/~vanschm/resources/uploads/dgb/database-documentation.pdf

Yamamoto, M. (1999). *Animacy and Reference: A Cognitive Approach to Corpus Linguistics*. Amsterdam: John Benjamins.

Zarrieß, S. (2016). *Syntactic and referential choice in corpus-based generation: modeling source, context and interactions.* Ph.D. thesis,University of Stuttgart, Stuttgart.

Zarrieß, S., and Kuhn, J. (2013). "Combining referring expression generation and surface realization: A corpus-based investigation of architectures," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia (Stroudsburg, PA: Association for Computational Linguistics), 1547–1557.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer CL and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

frontiers in Psychology

ORIGINAL RESEARCH
published: 01 March 2016
doi: 10.3389/fpsyg.2016.00214

Semantic Relations Cause Interference in Spoken Language Comprehension When Using Repeated Definite References, Not Pronouns

Sara A. Peters[1, 2*], *Timothy W. Boiteau*[2] *and Amit Almor*[2, 3]

[1] *Social and Behavioral Sciences Department, Newberry College, Newberry, SC, USA,* [2] *Department of Psychology, University of South Carolina, Columbia, SC, USA,* [3] *Linguistics Department, University of South Carolina, Columbia, SC, USA*

The choice and processing of referential expressions depend on the referents' status within the discourse, such that pronouns are generally preferred over full repetitive references when the referent is salient. Here we report two visual-world experiments showing that: (1) in spoken language comprehension, this preference is reflected in delayed fixations to referents mentioned after repeated definite references compared with after pronouns; (2) repeated references are processed differently than new references; (3) long-term semantic memory representations affect the processing of pronouns and repeated names differently. Overall, these results support the role of semantic discourse representation in referential processing and reveal important details about how pronouns and full repeated references are processed in the context of these representations. The results suggest the need for modifications to current theoretical accounts of reference processing such as Discourse Prominence Theory and the Informational Load Hypothesis.

Keywords: reference, repeated name penalty, pronouns, semantic relations, spoken language comprehension

OPEN ACCESS

Edited by:
Roger P. G. Van Gompel,
University of Dundee, UK

Reviewed by:
Stanka A. Fitneva,
Queen's University, Canada
Alan Garnham,
University of Sussex, UK
Federica Cavicchio,
Università degli Studi di Trento, Italy

Correspondence:
Sara A. Peters
sara.ann.peters@gmail.com

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

Received: *12 March 2015*
Accepted: *03 February 2016*
Published: *01 March 2016*

Citation:
Peters SA, Boiteau TW and Almor A (2016) Semantic Relations Cause Interference in Spoken Language Comprehension When Using Repeated Definite References, Not Pronouns. Front. Psychol. 7:214. doi: 10.3389/fpsyg.2016.00214

INTRODUCTION

Coherent discourse can be established via different forms of repeated reference to the same referent, such as repeated names (e.g., *Jane*), definite descriptions (e.g., *the girl*), and pronouns (e.g., *she*, *her*). The form used for repeated references is generally related to the discourse status of the referent (Almor and Nair, 2007). For example, full definite descriptions are often used to introduce referents that were not previously mentioned or to refer to referents that were previously mentioned but are not currently salient in the discourse (Gundel et al., 1993). In contrast, pronouns are often and naturally used to refer to previously mentioned referents that are salient in the context of the discourse (Ariel, 1990; Gundel et al., 1993). One of the clearest empirical demonstrations of the relationship between referential form and referents' discourse status is the repeated name penalty (RNP). This effect was first demonstrated as the slower reading of repeated proper names than pronouns when referring to the most salient referent in the discourse (Gordon et al., 1993). The RNP was later extended to repeated definite references, which are read slower when the referent is focused than when it is not (Almor, 1999). Multiple theories have been developed to explain the relation between reference form and the discourse status of referents (for a review see Almor and Nair, 2007). Although many of these theories explain this relation on the basis of general language and memory mechanisms that are not modality specific, much of the relevant empirical findings are based on reading paradigms. One aim of the present work was therefore to test whether the RNP occurs in spoken language comprehension, and if so, to test the predictions of existing theories about the timing and presence of the memory processes that underlie this effect.

Gordon et al. (1993) originally explained the RNP in terms of Centering Theory (Grosz et al., 1995), which Gordon and Hendrick (1998) later developed into Discourse Prominence Theory (DPT). According to DPT, repeated full references are initially interpreted as introducing new discourse entities that later require integration with the representation of the previous discourse, while pronouns are initially interpreted as co-referential, and thus do not generate new representations. In addition, referential processing is governed by a set of construction rules applied as part of building and maintaining a discourse representation. When a proper name is encountered, a construction rule generates a new representation in the discourse model, but when a pronoun is used, a different construction rule searches for a matching referent in the list of previously mentioned referents in decreasing order of prominence (Gordon and Hendrick, 1998). This model explains the RNP as reflecting the application of a special "equivalence" construction rule that reconciles the representation of the new referent generated by the proper name construction rule, and the representation of the referent already in memory. The rule searches the list of referents in ascending prominence order, thus taking longer to identify matches with prominent referents than with non-prominent referents. DPT therefore attributes the RNP to the time needed to merge the newly generated representation evoked by repeated full references with the existing discourse representation. Importantly, DPT also argues that repeated references are processed similarly to new references, at least during the initial stages of processing.

An alternative approach is the Informational Load Hypothesis (ILH; Almor, 1999, 2000, 2004; Almor and Nair, 2007), which attributes the RNP to memory interference between the representation of the referential expression and the existing representation of the referent in memory. In its original formulation (Almor, 1999, 2004), the theory emphasized the interaction between pragmatic principles and memory constraints but did not explicitly address the detailed processing time course of referential expressions. However, a more recent version of the theory (Almor and Nair, 2007) includes specific claims about the stages that are involved in processing referential expressions. According to this version, referential processing involves multiple stages that are differentially affected by the salience of the referent in the discourse. In Stage 1, the incoming referential expression undergoes lexical processing before it can be integrated into the discourse representation. This stage results in the activation of a representation of the referential expression that is initially separate from the prior representation of the discourse. Priming from a salient referent may facilitate this initial activation. Stage 1 in this view is compatible with Ledoux et al.'s (2007) finding of an initial stage of processing involving priming due solely to repetition, regardless of other co-referential processes. In Stage 2, this newly activated representation is integrated with the prior discourse representation. These two stages can overlap, but Stage 2 generally takes longer to complete than Stage 1.

Although both DPT and the ILH argue that processing definite reference results in the formation of a new representation, the two theories differ in their view of how and when this representation is processed. In DPT, repeated reference is processed just like a new reference. According to this theory, the difference between a new reference and a repeated one only occurs when a potential referent is not found through the serial search of current referents. Thus, in this account, repeated, and new references are processed similarly. In contrast, according to the ILH, the newly formed representation can interact with the existing discourse representation right from the start. In particular, the initial formation of the reference (Stage 1) can be facilitated by semantic overlap with an existing representation, but the integration of the representation with the discourse (Stage 2) is prone to interference due to semantic overlap. According to the ILH, repeated references are simply more likely to trigger these effects than new references, which do not necessarily overlap semantically with existing representations (Almor and Nair, 2007).

According to the ILH, maintaining similar, yet distinct representations results in interference until integration of the new representation is complete. This interference reflects the effort associated with maintaining simultaneously activated representations in a limited-capacity memory system. Therefore, the extent of this interference is affected by the activation of the referent in memory, such that salient referents can cause more interference during integration than less salient ones. This interference is also affected by the semantic overlap between the representations of the referential expression and the existing representation of the referent in memory. When the referent is already salient in the discourse, high overlap between the two may result in greater interference. In this view, pronouns minimize memory interference during integration because they do not evoke a rich representation in memory. Pronouns are therefore generally preferred when the referent is salient and can be easily identified. Thus, according to both the ILH and DPT, repeated full reference evokes an initial representation that is separate from the existing representation of the referent in memory. However, while DPT considers this representation to be equivalent to that of a new reference, the ILH considers it as an initial lexical representation that could lead to memory interference during integrative processing. Specifically, Almor (1999) argued that a high level of semantic similarity between a referential expression and the memory representations of previously mentioned referents makes it harder to maintain the representation of the discourse in a working memory mechanism specializing in the manipulation of semantic information. Almor based this argument on an analogy to the detrimental effect phonological similarity has on word list recall (Baddeley, 1992, 1996). This analogy assumes that the representation of discourse referents relies on a limited capacity memory system of semantic representations and that referential processing activates and manipulates these representations.

Cowles et al. (2010) presented evidence against this analogy. Using a word recall task modeled after the original Baddeley experiments, Cowles et al. replicated Baddeley's original finding of memory interference associated with phonological overlap between words in the memory list, but found no evidence for similar interference associated with semantic overlap. This suggests that the difficulty associated with semantic overlap during reference processing may reflect a memory process dissimilar from phonological interference in Baddeley's recall task. One phenomenon that has been linked with detrimental effects of semantic overlap on memory recall is the Cue-Overload effect (Watkins and Watkins, 1975), in which the efficiency of a retrieval cue is reduced when it has been used previously as a retrieval cue for a semantically similar but different item set. This likely reflects the activation of superfluous information, rendering the intended recall target less distinctive and therefore interfering with selection and processing.

A similar phenomenon may occur during reference processing and may help explain why semantic overlap can hinder the processing of referential expressions. In particular, it may be that a high level of semantic overlap between a referential expression and the existing representation of a referent increases the activation of the referent and the information already associated with it in long-term memory. This spreading activation occurs at the expense of processing other information in the discourse. Thus, semantic overlap and repetition initially result in facilitation; however, this facilitation results in the activation of other information in memory, and consequently in a reduced ability to process new discourse information until the integration of the referential expression is complete. This explanation is in fact more compatible with the vast literature on the facilitative memory effects of semantic overlap than the original explanation made in Almor (1999).

These theories appeal to general language and memory processes that are not specific to reading. However, most of the relevant research on the RNP has employed reading-based paradigms. In a study of reference processing in spoken language comprehension, Dahan et al. (2002) showed that when the referent is salient but not in focus, repeated definite references can be used felicitously without penalty. However, Dahan et al. only used stimuli in which the target references appeared in the grammatical object position of imperative instructions (e.g., "Put the X above the Y"). Thus, the study does not answer the question of whether the RNP occurs in spoken language comprehension. Another study that examined the RNP in spoken language comprehension is Almor and Eimas (2008), who found an initial facilitation in a lexical decision task but poor memory in a delayed recall task of information from discourses with repeated definite references to salient referents. Although this result suggests that the RNP extends to spoken language comprehension, it provides only limited information about the time course of the underlying processing (i.e., initial facilitation followed by delayed interference) or about the specifics of the underlying memory processes.

Given this previous work, the goals of the present study were to investigate the RNP in spoken language comprehension, examine the time course of reference processing in spoken language comprehension, and better understand the underlying memory processes and their influences. Experiment 1 examined whether an effect analogous to the RNP can be observed in spoken language comprehension using the visual world paradigm (VWP; Tanenhaus et al., 1995) and at the same time examine the time course and nature of the underlying memory processes in order to test the theories presented above. The visual displays used in the experiments were borrowed from Yee and Sedivy (2006) and contained items that were semantically related to one another within the displays. While these semantic relations were not exclusively based on shared category membership as in previous work on semantic overlap in reference processing (e.g., Almor, 1999), the inclusion of the semantically related items within Experiment 1 aimed to provide a more general test of activation of semantic information during reference processing and the RNP. Thus, this experiment aimed to test the following predictions of DTP and ILH in regards to reference processing and interference:

1. According to DPT (Gordon and Hendrick, 1998), repeated definite references are initially processed like new references. The similarity between the two forms is greatest when the referent is salient because the search for an existing referent in the discourse representation in the case of a full noun reference proceeds in increasing order of salience Thus, the referent search for a repeated definite reference to the most salient referent will require scanning the entire list of potential referents just as for a new definite reference.
2. According to the ILH (Almor, 1999), both pronouns and repeated references are initially interpreted as referring to a previously mentioned and salient referent. Although the early processing of repeated references could be facilitated due to repetition priming, memory interference during integrative processing should lead to delayed processing later on, especially if the reference was to a highly salient referent.
3. The two possible views of the underlying interference that were reviewed can also be tested in Experiment 1. If, in line with Almor (1999), semantic competitors within the display create interference, the resulting activation should resemble the semantic competitor effect observed by Yee and Sedivy (2006) during initial lexical processing. In contrast, interference that is generated by items that have been linked earlier in discourse with a salient referent, should be reflected in the activation of the information that was originally mentioned. This outcome would support the view that the interference underlying the RNP reflects the activation of information related to the referent at the expense of further processing the discourse.

Experiment 2 further tested the effect of explicitly mentioning a semantically related referent (e.g., *cat*) in previous discourse on the processing of a subsequent reference to a new target reference (e.g., *mouse*). In particular, this experiment aimed to determine whether such mention would have a facilitative or detrimental effect on processing a repeated reference to the target item later within the discourse. Specific predictions tested were as follows:

1. According to a cue-based retrieval explanation of reference processing, pre-existing semantic relations should facilitate the retrieval of the representation of the referent. If those referents have long term associations (semantic associations), the effect should be larger than if the association only was established in the discourse.
2. According to the ILH, the relationships (pre-existing or established in the discourse) should have a different effect on the processing of potential referents when repeated and pronoun references are used. The interference in processing repeated anaphors is expected to be greater with long-term relations in place between the referents than when the relationship between the referents has only been established in the discourse. This is hypothesized to be due to the increased activation of a related previously mentioned item than an unrelated previously mentioned item.

EXPERIMENT 1

We used the VWP in order to obtain detailed information about the time course of activation of potential referents. This paradigm has been previously used to study referential processing (e.g., Dahan et al., 2002) but not for examining the RNP. Listeners' gazes in the VWP are known to be closely time locked to the unfolding language input (e.g., Allopenna et al., 1998; Arnold et al., 2004). Additionally, while previous research has shown that characteristics of the display influence fixation patterns, here we control properties of the referents in the visual display and allow for a preview time before each item begins to minimize these effects, as explained further in the Materials Section. Therefore, although the theories we test here were not previously tested using the VWP, we assume, following other research on the VWP, that the effects of language on attention and eye movements are immediate and reliable. Furthermore, we assume that the eye movements reflect the processing of and attention to the *specific* on-screen referent currently fixated (Tanenhaus et al., 1995). Thus, if interference in reference resolution is caused by working memory processes, those delays should also be reflected in the eye-movement record, and the delay can be evaluated in the models used to compare fixations.

We constructed 3-sentence discourses that made a target referent salient in two ways: the target referent was mentioned in the grammatical subject position of two sentences prior to the critical sentence (Sentence 3), and it was referred to with a pronoun in the second of these sentences. Both manipulations have been previously shown to effectively increase salience and lead to the RNP (Gordon et al., 1993). In the critical sentences, we contrasted repeated definite references (the *Repeated* condition), pronoun references (the *Pronoun* condition), and definite references to new referents (the *New* condition). A separate study confirmed that these items elicit the RNP (Peters and Almor, 2006) in a text-based, self-paced reading paradigm that also served to pilot the items for the current work. In a critical experiment in that study, participants read discourses that were almost identical to the present items and took longer to read the third sentence in the Repeated condition than in the Pronoun condition. Another experiment in the study using the same items but without the second sentence, found that participants took longer to read the final sentence in the Pronoun condition than the Repeated condition. The second experiment thus showed that the RNP found in the first experiment was not simply a result of baseline differences in reading times between the critical sentences in the different conditions. In the present study, we contrasted for each of the referential conditions (Pronoun, Repeated, and New) fixations to the target referent (*Target*), a potential referent that was semantically related to the target referent (*Semantic Distractor*), a referent that was not semantically related to the target but that was mentioned in Sentence 1 (*Sentence 1-mentioned*), and an unrelated referent that was mentioned for the first time in the critical sentence (*Sentence 3-mentioned*). **Figure 1** shows a sample experimental display and **Table 1A** shows the corresponding verbal stimulus in all three conditions. Detailed predictions derived from the theories presented in the introduction are stated below, in terms of fixations to the visual displays based on the discourses used in the experiment. An important aspect of this design is that since our critical sentences occurred several seconds and two sentences after the pictures were originally presented, eye movement patterns are unlikely to reflect any effects of the visual properties of the depicted objects (Henderson and Ferreira, 2004).

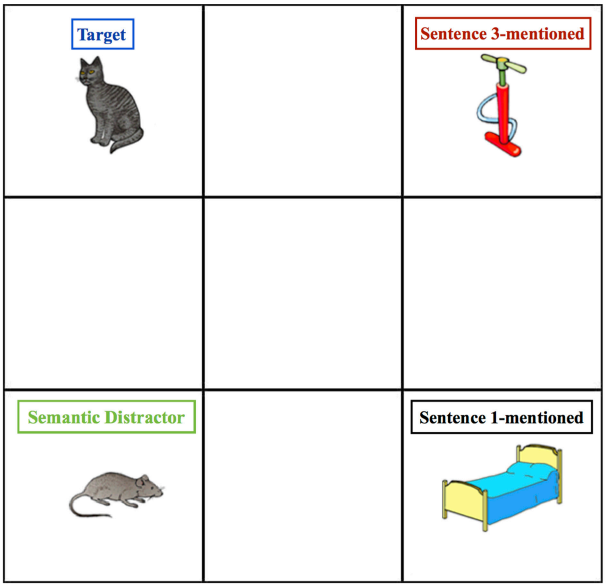


FIGURE 1 | Sample visual display, labeled. Colors correspond to data graphed in future figures.

1. DPT argues that repeated references are initially processed as new references. Therefore, in the critical sentence (Sentence 3), listeners should initially interpret both the repeated and the new references as mentions of a new object, and look at an object not yet mentioned, which is the Semantic

TABLE 1 | Sample verbal item in all three conditions from (A) Experiment 1 and (B) Experiment 2.

Sentence #	Condition	Sentence
(A)		
1		The cat is diagonal to the bed.
2		It is in the upper left corner.
3	*Pronoun*	It is next to the pump.
	Repeated	The cat is next to the pump.
	New	The pump is next to the cat.
(B)		
1	*Unrelated*	The cat is diagonal to the bed.
	Related	The cat is above the mouse.
2		It is in the upper left corner.
3	*Pronoun*	It is next to the pump.
	Repeated	The cat is next to the pump.

Sentence 1 and 3 varied by condition, Sentence 2 remained constant. Participants heard one combination of each item, and each participant heard 6 of each combination of Unrelated-Pronoun, Unrelated-Repeated, Related-Pronoun, Related-Repeated within the experiment. Sentences 1 and 2 were the same in all conditions, while Sentence 3 varied by condition. (B) Sample verbal item in Experiment 2 in all conditions.

Distractor (*mouse*) in both conditions. This should result in comparable number and rate of increase in looks to the Semantic Distractor in the New and Repeated conditions following the critical reference. However, this should not be the case in the Pronoun condition, because identifying the referent of a pronoun proceeds in decreasing salience order, thus leading to the immediate identification of the Target (*cat*) as the intended referent.

2. According to the ILH, the initial processing of repeated references could be facilitated but the Repeated condition should lead to interference later in the critical sentence (Sentence 3) as integration takes place and the discourse unfolds. This should be reflected in an overall smaller number and a lower rate of increase in looks to the second referent mentioned (Sentence 3-mentioned, *pump*) in the critical sentence in the Repeated condition relative to the Pronoun condition. As the effect of the interference dissipates, this rate of fixation is likely to increase resulting in a larger quadratic component in the Repeated relative to the Pronoun condition.
3. In terms of interference types, more gazes, or a higher rate of fixations to the Semantic Distractor than to the Sentence 3-mentioned before it is heard, will indicate the activation of semantic representations. Such activation would be similar to that generated by initial lexical processing, in line with Almor (1999). In contrast, more gazes and a higher rate of fixations to the previously mentioned referent (Sentence 1-mentioned, *bed*) than to another referent that has not been mentioned (Semantic Distractor, Sentence 3-mentioned) would suggest that the interference is related to the activation of the information that was originally mentioned with the Target referent. This would support the cue-overload view that the interference underlying the RNP reflects the activation of information related to the referent at the expense of further processing the discourse.

As previously noted, the VWP provides finely tuned time course information about the activation and processing of the information presented in the display. In order to fully utilize this information, we modeled the fixation results using growth curve analyses (GCA) (Mirman et al., 2008). As we explain below, GCA allows us to test predictions about both the number and rate of change of looks in the different conditions. Importantly, this type of analysis allows us to explore the dynamics of fixation changes as well as sustained attention to a referent in a time window of interest. This would have been impossible under the common approach of averaging fixations in separate (and typically large) time windows. To the best of our knowledge, the use of GCA to study these dynamics of reference processing is novel. To reduce concerns related to the dependence between fixations to the different pictures, the majority of our analyses compared fixations to a particular display item across discourse conditions (e.g., fixations to Semantic Distractor in Pronoun vs. Repeated conditions). However, when the critical theoretical prediction rests on the difference in looks to different display items, we also included an analysis comparing fixations to different display items.

METHOD

Participants

Forty-nine undergraduate students recruited from the University of South Carolina Psychology Department's participant pool participated in this experiment for course credit. All participants provided informed consent in accordance with the University's IRB. All participants were native speakers of American English.

Materials

As shown in **Figure 1** and **Table 1A**, each item consisted of a pictorial display showing four objects arranged in the corners of a 3 × 3 grid, and a corresponding 3-sentence discourse. The pictorial displays were taken from Yee and Sedivy (2006) and included 24 experimental, 48 filler, and 4 practice displays. Experimental displays showed 2 semantically related objects [e.g., a Target (*cat*) and a Semantic Distractor (*mouse*)] and 2 objects whose names matched the names of the semantically related objects for word frequency. Yee and Sedivy (2006) also validated the items using a picture naming task to verify the pictures evoked the intended linguistic label and to ensure that the control items did not compete phonologically or semantically with other items on the display (see Yee and Sedivy, 2006, for a complete description of these pictures). The items were randomly placed in the grids in static positions a priori, which resulted in the same grid being viewed by each participant. However, the objects (target, semantic distractor, and their controls) were equally distributed in the grids across the experiment.

The 3-sentence discourses described the location of the objects in the grid and were played one after the other with a 600 ms delay between sentences starting after the display was shown for 2 s. Participants were given these 2 s in order to familiarize themselves with the objects and their location in the grid before hearing the auditory description, which they had to verify. Sentence 1 (∼2454 ms) began by describing the location of the Target referent (one of the semantically related objects) in relation to an unrelated object (Sentence 1-mentioned) using definite references. Sentence 2 (∼2294 ms) always referred to the Target using a pronoun and described the absolute position of the Target in the grid without referring to any other object. Sentence 3 varied by condition (*Pronoun* ∼2025, *Repeated* ∼2583 ms, *New* ∼2595 ms), and was the critical sentence used for analysis.

Verbal stimuli were recorded by a native female speaker of American English (S.A.P.) and were edited using Adobe sound editing software. All the experimental items included the same recorded version of Sentences 1 and 2. Sentence 3 was recorded separately for each condition. Items were presented in a random order that was different for each participant. Each participant was presented with each experimental item once, such that they responded to eight items in each condition. Across all participants, each item appeared in each condition a similar number of times. Experimental items were always true, and 12 of the 48 fillers were also true, such that overall the verbal descriptions in exactly half of the trials were true. The false statement in the false filler items appeared exactly 12 times in each sentence position (1, 2, and 3).

Apparatus

Participants' eye movements were recorded using a stationary chin rest ASL 6000 which sampled eye position at 240 Hz. Visual stimuli were presented on a 19" Dell CRT monitor positioned 62 cm directly in front of participants. The experiment was controlled by a Dell computer running the E-prime software (Schneider et al., 2002).

Procedure

The experiment began with four practice trials, before which the experimenter calibrated the eye tracker using a 9-point calibration procedure. Calibration was repeated every four trials during the experiment if needed. This was done only when participants needed to exit the tracker momentarily for a break, or when visual analysis of drift indicated that participants had shifted head position. Participants were told that they would be looking at pictorial displays and hearing short discourses. They were instructed to listen to the discourses and decide whether they accurately described the displays. At the end of the recorded discourse, the pictorial display was replaced with a screen instructing participants to indicate their response by clicking on the words *True* or *False*. Participants were informed that even one false statement in the discourse made the entire discourse false. When participants clicked on a choice, they were given immediate feedback on their accuracy. Response accuracy was recorded and analyzed to ensure that participants were performing the task. The data from four participants whose accuracy was lower than 3 standard deviations below the median accuracy of all participants were removed from the analysis. In addition, 3 participants were also removed because proper calibration was not maintained throughout the course of the experiment. The data from the remaining 42 participants are reported below.

Eye Tracking Data Analysis

Raw eye position data from each participant were transformed into fixations using ASL Results 2.0 analysis software, following the procedure recommended by ASL. Fixations were then matched with the E-prime data file of the participant in order to determine the sentence, item, condition, and object fixated for each fixation. Proportions of fixations to each display item type were then calculated for each subject in each condition in consecutive 25 ms time windows by averaging across the experimental items the participant saw in each condition. These proportions of fixations to display items were the dependent variables for the analyses described below.

GCAs

Because our focus was on the time course of processing the different reference form we chose to employ GCA of the fixation proportion data, following the procedure outlined in Mirman et al. (2008) and Mirman (2014). Unlike traditional analyses of variance, GCA includes time as a predictor in the model and analyzes the effect of the different conditions on the rate of change in fixation proportions. To ensure that our fixation proportion results were not biased by the relatively small number of observations per condition we repeated our analyses using the empirical logit transformations. As the results were identical and only served to increase power and fit, we only report the more readable proportion of fixations results.

Our analyses aimed to test the predictions of the different theories and focused on rapid changes in a short, 500 ms time window following the critical events in the sound files. For testing the lexical semantic competitor effect, in keeping with previous research (e.g., Yee and Sedivy, 2006), the window of analysis was 500 ms, starting 100 ms before the offset of the reference. This start point was maintained to capture the effect, as it has been shown to disappear rapidly when the competitor does not receive further mention. We also utilized a 500 ms time window for testing the consideration of possible referents after hearing the target reference; in order to better compare the effect of hearing pronouns, which are short, to the effect of hearing longer definite references, the window starting at the offset of the critical reference. Thus, the choice of this relatively short time window was based on our focus on the processes that occur immediately after listeners process the critical reference and minimize as much as possible the effect of inter-item differences in the subsequent linguistic input. While these analyses have been previously used in studies of single word processing (e.g., Chen and Mirman, 2015) and shifts in attention during conversation (Boiteau et al., 2014), they have not been used specifically in the type of reference processing comparison we present here. We chose to use these analyses as they allowed us to look closely at the rapid changes in processing and attention that are predicted by some of the contrasted hypotheses. While this short time window can only include one or two fixations, the high sampling rate and the averaging of items and subjects resulted in a rich data set that adequately reflected changes to the sustained attention to the different objects. Indeed, if these data were to be compared using a simple ANOVA, we would lose information regarding the finer differences in the processing of the different anaphoric expressions. An example of this type of finely tuned information is the average rate at which participants switch from looking from one display item to another. This rate of change can reveal the strength of the attentional commitment to a referent of an anaphoric expression just heard, which can be directly modeled as the quadratic component of a GCA. In contrast, this rate of change is not directly expressed in a traditional ANOVA, unless the means of consecutive *post-hoc* sized time windows are compared, with a likely loss of statistical power.

In our GCA, the models contained two levels, the first of which (Level-1, see top line in Equation 1), captures the effect of time on the performance of participant *i* in condition *j*:

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i}{}^{*}\ Time_{ij} + \beta_{2i}{}^{*}\ Time^{2ij} + \beta_{3i}{}^{*}\ Time^{3ij} + \varepsilon_{ij} \\ \beta_{0i} &= \gamma_{00} + \gamma_{01}{}^{*}\ \text{Condition} + \zeta_{0i} \\ \beta_{1i} &= \gamma_{10} + \gamma_{11}{}^{*}\ \text{Condition} + \zeta_{1i} \\ \beta_{2i} &= \gamma_{20} + \gamma_{21}{}^{*}\ \text{Condition} \\ \beta_{3i} &= \gamma_{30} + \gamma_{31}{}^{*}\ \text{Condition} \end{aligned} \tag{1}$$

In these models, the first order (linear/slope) effects of time (Time) reflect the overall rate of fixation change while second order (quadratic) effects (Time^2) reflect the rise and fall of the change in fixation rate, and third order (cubic) effects (Time^3) reflect higher order changes in the change rate of fixation rates (Mirman et al., 2008). Since we were interested in fixations on target objects over short time windows, which included non-linear change trends but not ones that were highly complex, all the models we tested included fixed effects of time up to the third power, as shown in Equation 1, although when less complex models were identified as having better fit, they were chosen.The second level in GCA is used to estimate the effect of condition on the intercept (γ_{01} in the second line of Equation 1) and on the time course at the different orders (γ_{k1} in the lines 3–5 in Equation 1) by adjusting for individuals and conditions. Our models always included a random effect of participants on the intercept (ζ_{0i} in line 2) and slope (ζ_{1i} in line 3), thus allowing both the estimated baseline fixation proportion and rate of change in fixations to vary across individuals, which serves to measure of the variability across participants within the model.

In GCA, the effect of the condition is inferred by its necessity for the fit of the model in a process of model comparison. The best-fitting model is chosen according to a criterion that optimizes model fit and number of degrees of freedom, such that the simplest model that fits the data no worse than more complex models is chosen. Here, again following Mirman et al. (2008), complexity of models varied by the order of the time coefficient included in the model. Within our analyses, the time variables were represented by orthogonal, mean centered polynomials in order to eliminate the possible confounding effects of multicollinearity. Note that due to the centering of the time variables, intercept coefficients represent the middle time point in the analyzed time range (e.g., 250 ms from the start of the time window) and not the first time point.

Despite its advantages in analyzing change, as in any analysis that is based on model comparison, the use of GCA carries a risk of over fitting the data. To reduce this risk, statistical texts recommend that the results of GCA are interpreted by considering the terms included in the chosen model, the parameter estimates within the chosen model, and the visual inspection of the fitted model (Long, 2012). Following this advice and in order to help readers interpret our results and analyses we provide the details of all three for each analysis. The selected models' parameter estimates are shown in tables within the paper. These tables also present *p*-values of individual coefficients, which we calculated following Mirman (2014), using a Unit Normal Table approximation to the critical *t*-values. Together with these tables we also provide figures with the best-fitting growth curve model overlaid on the mean observed. To facilitate the flow of presentation in the paper, the process of model selection including the differences in the fit of the contrasted models, are documented in the table section of the Supplemental Materials. Thus, the analyses in Supplemental Materials correspond with the tables included within the results, and offer the interested reader additional model fit information.

We should emphasize that our main goal is to compare the time course of reference processing through the sustained attention to referents in the different conditions. Therefore, our analyses focus on whether the best-fitting model includes any interaction effects involving condition and the intercept, or condition and any of the time terms.

Analyses were carried out using the R statistical software package (v.3.1.1, R Core Team, 2014), and the lme4 (Bates et al., 2014) and lmerTest (Kuznetsova et al., 2014) software packages, which run mixed-effects models including the GCAs used here. In order to avoid intractably complex analyses with high order interactions, we performed a series of planned analyses on subsets of the data aiming to test the specific theoretical predictions outlined above.

RESULTS

Eye tracking data from 42 participants were preprocessed and analyzed as described above. We present below the analyses of the fixations during Sentences 1 and 3.

Sentence 1

Sentence 1 fixations were analyzed to replicate Yee and Sedivy's semantic competitor effect (2006). The current study utilized references to target objects in non-imperative, descriptive sentences. As Yee and Sedivy found that listeners looked at the Semantic Distractor after the offset of the target word, we carried out a GCA contrasting the fixations to the Semantic Distractor to fixations to Sentence 1-mentioned and to Sentence 3-mentioned during a 500 ms time window starting 100 ms before the offset of the Target referent. The participant had not yet heard the name of the three other items, and so any advantage for the Semantic Distractor (*mouse*) over the other two items (*bed* and *pump*) during this period would necessarily reflect a semantic competitor effect. As in all the analyses in this paper, all the contrasted models included level-1 terms corresponding to intercept, slope, and linear, quadratic, and cubic terms for Time. Here the model also included terms representing the destination of the fixation [Semantic Distractor (*mouse*),

TABLE 2 | Coefficient estimates in the best-fitting model, (slope) for proportion of fixations to the Semantic Distractor, Sentence 1-mentioned (S1-M), and Sentence 3-mentioned (S3-M) in Sentence 1 in a 500 ms time window starting 100 ms before the offset of the Target.

Coefficient	*Est.*	*Std. Error*	*t*	*p* <
Intercept	0.1837	0.0061	30.173	0.001
Time	−0.1030	0.0150	−6.884	0.001
S1-M	−0.0503	0.0034	−14.596	0.001
S3-M	−0.0642	0.0034	−18.624	0.001
Time*S1-M	−0.0385	0.0154	−2.494	0.01
Time*S3-M	−0.0822	0.0154	−5.334	0.001

Proportion of fixations to the Semantic Distractor provided the baseline group.

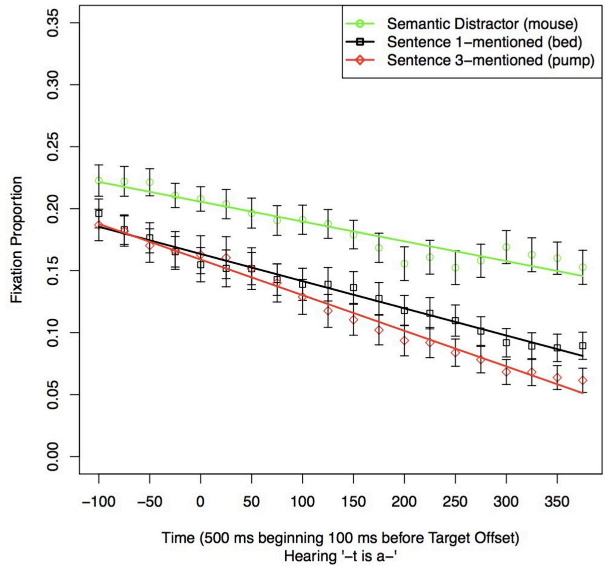


FIGURE 2 | Proportion of fixations to other items in the display in the 500 ms time window starting 100 ms before the offset of the reference to the Target in Sentence 1. Error bars indicate the standard error of the mean for the condition across subjects in the time-window.

Sentence 1-mentioned (*bed*), or Sentence 3-mentioned (*pump*)], and their interaction with the various time terms.

The coefficient estimates of the best-chosen model are given in **Table 2** (see the correspondingly numbered tables in Supplemental Materials for the model comparison leading to the model's choice). **Figure 2** shows the proportion of fixations to these three pictures as well as the fit estimate lines for the best-fitting model. For the remainder of our analyses, we will list these components in the same order. In the chosen model, the destination of fixation only affected the intercept and the linear time term, but not any of the higher order time terms.

This analysis shows that, at the offset of the Target (*The cat*), listeners looked reliably more often and were slower to look away from the Semantic Distractor (*mouse*) than the other two objects that have not been mentioned yet (*bed* and *pump*). This analysis thus confirms that the semantic competitor effect observed by Yee and Sedivy (2006) occurs in declarative sentences like the ones used here.

Sentence 3

Prediction 1

In order to test DPT's prediction that repeated references are processed like new references, we looked for differences between the Repeated and New conditions. Because these two conditions differed in whether the first reference in the critical third sentence was to the Target (Repeated condition, *The cat…*) or to Sentence 3-mentioned (New condition, *The pump…*), we only looked at fixations to pictures of the other two items: the Semantic Distractor (*mouse*), which was not mentioned in any of the conditions, and Sentence 1-mentioned (*bed*), which was mentioned in the first sentence in all the conditions. This ensured that any differences in fixations between the Repeated and New conditions do not merely reflect a baseline difference between looks to a picture that was mentioned before vs. one that was not. According to DPT, both the Repeated and New conditions should be similar and both should differ from the Pronoun condition.

Our first analysis contrasted the proportion of fixations to the Semantic Distractor (*mouse*) in the three conditions in the 500 ms time window following the offset of the critical first reference in Sentence 3. This time window was chosen because it captures mainly the effects of processing the first reference on eye movements and our focus in this analysis was on whether the initial and immediate processing of new and repeated references is similar. In all conditions, the Semantic Distractor has not been mentioned. If repeated and new references are processed similarly, then there should be no differences between the two conditions in fixations to an object that had still not been mentioned and both these conditions should differ from the Pronoun condition.

The results of the analyses are shown in **Table 3A** and **Figure 3**. As is shown, participants looked more at the Semantic Distractor in the New condition than in either the Repeated condition or the Pronoun condition in the first 250 ms, even though it was not mentioned in any of these conditions. The best-fitting model was cubic. This analysis illustrates the importance of using GCA, as averaging fixations over the time window would have likely missed this effect, and would not have provided any information about the dynamics of the fixations.

Our next GCA contrasted the proportion of fixations to Sentence 1-mentioned (*bed*) in the three conditions in the same time window. In all conditions, Sentence 1-mentioned was mentioned together with the Target (*cat*) in Sentence 1. If repeated and new references are processed similarly, then there should be no differences between the two conditions in fixations to an object that was previously mentioned.

The results of the analyses are shown in **Table 3B** and **Figure 4**. As is shown in both table and figure, there was an intercept effect reflecting more looks at Sentence 1-mentioned (*bed*) in the Repeated condition than in either the New or Pronoun conditions. There was also an effect of condition on slope reflecting a contrast between the steady increase in fixations to the Sentence 1-mentioned in the Repeated condition across the time window, in comparison to the steady decrease of looks in

TABLE 3 | Coefficient estimates for the best-fitting models in the 500 ms time window starting at the offset of the critical reference in Sentence 3 to the (A) Semantic Distractor (cubic) and (B) Sentence 1-mentioned (slope) in the Pronoun, Repeated and New conditions.

Coefficient	*Est.*	*Std. Error*	*t*	*p* <
(A) SEMANTIC DISTRACTOR				
Intercept	0.0853	0.0088	9.744	0.001
Time	0.0045	0.0225	0.200	*n.s.*
Time^2	0.0245	0.0159	1.542	*n.s.*
Time^3	0.0205	0.0159	1.287	*n.s.*
Pronoun	0.0181	0.0050	3.600	0.001
New	0.0350	0.0050	6.963	0.001
Time*Pronoun	−0.0193	0.0225	0.860	*n.s.*
Time*New	−0.1645	0.0225	−7.317	0.001
Time^2*Pronoun	−0.0399	0.0225	−1.773	0.08
Time^2*New	−0.0158	0.0225	−0.705	*n.s.*
Time^3*Pronoun	−0.0319	0.0225	−1.418	*n.s.*
Time^3*New	0.0363	0.0225	1.616	0.11
(B) SENTENCE 1-MENTIONED				
Intercept	0.1352	0.0128	10.558	0.001
Time	0.0572	0.0262	2.182	0.05
Pronoun	−0.0373	0.0050	−7.380	0.001
New	−0.0621	0.0050	−12.297	0.001
Time*Pronoun	−0.0501	0.0226	−2.218	0.05
Time*New	−0.1581	0.0226	−7.000	0.001

Proportion of fixations in the Repeated condition provided the baseline group.

the New condition and the barely unchanged rate of looks in the Pronoun condition.

Together, the two analyses indicate that listeners process repeated and new references quite differently: they consider other referents from the previous sentence after hearing a repeated reference and consider new referents besides the one mentioned, when hearing a reference to a previously unmentioned referent. Thus, overall, in contrast to the predictions of DPT, repeated and new definite references were not processed similarly, in that repeated definite references increased fixations to a referent that was previously mentioned (Sentence 1-mentioned, *bed*), but new definite references did not.

Prediction 2

To test the predictions of the ILH that pronouns and repeated references are interpreted as referring to the target but repeated names lead to interference, our remaining analyses focused on differences between the Pronoun and Repeated conditions. The New condition was not included because the Target referent was not mentioned first. For clarity, we start with a separate analysis of fixations toward each of the 4 displayed objects in the Pronoun vs. Repeated conditions.

Target (cat)

First, in order to test whether looks to the target differed following pronoun and repeated references, we analyzed fixations to the Target (*cat*) in the Pronoun and Repeated conditions. The results of the analyses are shown in **Table 4A** and **Figure 5A**. As is shown in both table and figure, participants looked more often at the Target in the Pronoun condition than in the Repeated condition, but this tendency did not change across the 500 ms time window. Thus, while there were more looks to the Target in the Pronoun than in the Repeated condition, the rate of looking away from the Target as Sentence 3 unfolded was comparable in the two conditions.

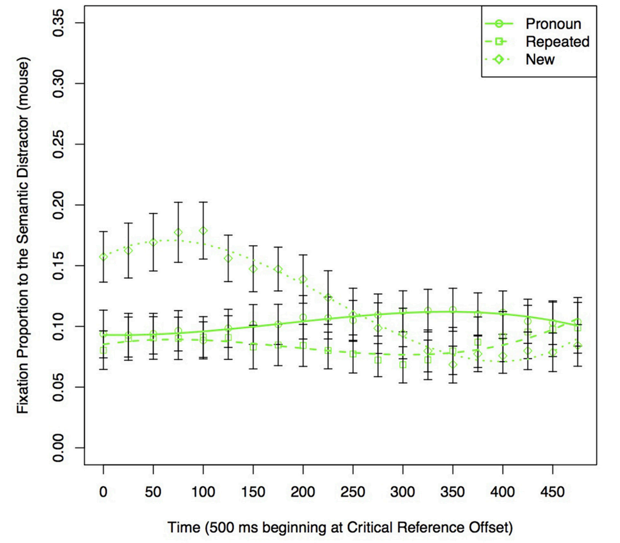


FIGURE 3 | Proportion of fixations to the Semantic Distractor picture (e.g., "the mouse") in the display in Sentence 3 in the 500 ms time window after the offset of the critical reference ("the cat"). Fixations are graphed by condition: either Pronoun, Repeated, or New.

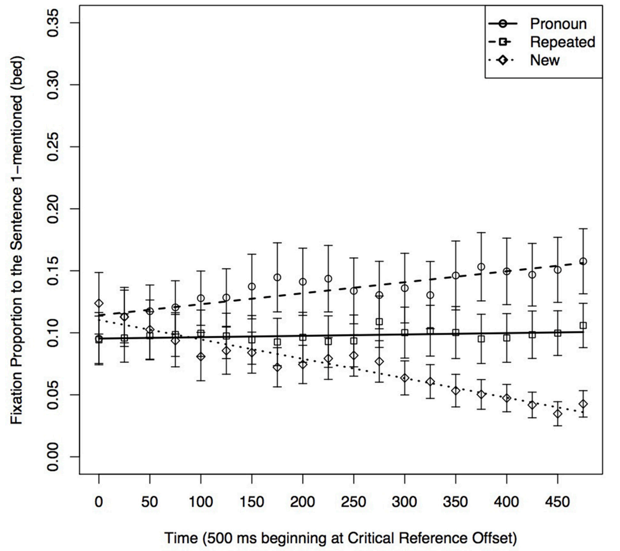


FIGURE 4 | Proportion of fixations to the Sentence 1-mentioned picture (bed) in the display in Sentence 3 in a 500 ms time window after the offset of the critical reference. Fixations are graphed by condition, Pronoun, Repeated, or New.

TABLE 4 | Coefficient estimates in the best-fitting models of proportion of fixations to display items in the Pronoun and Repeated conditions in a 500 ms time window starting at the offset of the critical reference in Sentence 3: (A) the Target (cat; Model 1); (B) Sentence 3-mentioned (pump; Model 2); (C) Semantic Distractor (mouse; Model 3); (D) Sentence 1-mentioned (bed; Model 4).

Coefficient	*Est.*	*Std. Error*	*t*	*p* <
(A) MODEL 1. TARGET				
Intercept	0.6576	0.0297	22.670	0.001
Time	−0.0971	0.0389	−2.498	0.05
Repeated	−0.0367	0.0075	−4.888	0.001
(B) MODEL 2. SENTENCE 3-MENTIONED				
Intercept	0.0927	0.0134	6.943	0.001
Time	0.0486	0.0229	2.120	0.05
Repeated	−0.0105	0.0043	−2.448	0.02
(C) MODEL 3. SEMANTIC DISTRACTOR				
Intercept	0.1034	0.0115	8.960	0.001
Time	0.00142	0.0217	0.654	*n.s.*
Repeated	−0.181	0.0040	−4.485	0.001
(D) MODEL 4. SENTENCE 1-MENTIONED				
Intercept	0.0980	0.0156	6.269	0.001
Time	0.0007	0.0285	0.249	*n.s.*
Repeated	0.0373	0.0049	7.566	0.001
Time*Repeated	0.0501	0.0220	2.274	0.05

Proportion of fixations in the Pronoun condition provided the baseline group.

Sentence 3-mentioned (pump)

Second, in order to test whether the two reference types led to differences during the processing of the remainder of Sentence 3, we analyzed fixations to the picture of the item that was mentioned second in this sentence (Sentence 3-mentioned, *pump*). The results of the analyses are shown in **Table 4B** and **Figure 5B**. The chosen model included an effect of condition on the intercept. As is shown in both the table and figure, the intercept effect was due to participants looking at the second-mentioned entity in the critical sentence more often in the Pronoun than in the Repeated condition.

Overall, the first two analyses show that (1) participants looked less often at the Target (*cat*) in the Repeated condition than in the Pronoun condition, and (2) participants looked more often at Sentence 3-mentioned (*pump*) in the Pronoun than the Repeated condition. Together, we interpret these effects as showing that pronouns were associated with quicker processing of the target as well as quicker processing of the second mentioned referent in the sentence. In other words, the pronoun condition showed less interference than the repeated condition.

Semantic distractor (mouse)

Next, we reanalyzed fixations to the Semantic Distractor without the New condition. The results of the analyses are shown in **Table 4C** and **Figure 5C**. The chosen model included an intercept effect of condition. The intercept effect indicated a greater number of fixations to the Semantic Distractor in the Pronoun than in the Repeated condition.

Sentence 1-mentioned (bed)

We also reanalyzed looks to Sentence 1-mentioned with only the Pronoun and Repeated conditions in the 500 ms time window. The results of the analyses are shown in **Table 4D** and **Figure 5D**. This model included effects of condition on the linear Time component indicating that participants looked more often at Sentence 1-mentioned in the Repeated condition than in the Pronoun condition throughout the 500 ms following the offset of the Target, and the difference increased toward the end of the time window. We interpret this finding as an indication of a greater activation of the previously mentioned referent in the Repeated condition than in the Pronoun condition. The fact that this effect increased over time indicates that this interference became more pronounced as processing progressed.

Prediction 3

The analysis above (Sentence 1-mentioned, *bed*) also tests Prediction 3. This prediction suggested that in the Pronoun and Repeated conditions, after hearing the Target, there should be fewer looks to the Sentence 1-mentioned than to other items that have not been mentioned. There are instead more looks to this item, particularly in the Pronoun condition.

DISCUSSION

Our data are not compatible with the DPT prediction of similar processing of new and repeated references. In our experiment, the New and Repeated conditions led to distinctively different fixation patterns to both the Semantic Distractor (*mouse*) and Sentence 1-mentioned (*bed*). The Semantic Distractor was not mentioned previously and, according to DPT, should have been considered a good "new" referent in both conditions. Sentence 1-mentioned was mentioned in the previous discourse, and, according to DPT, should not have been considered a good "new" referent in both conditions. Critically, this comparison did not involve looks to the Target which was a previously mentioned item in the Repeated condition and an unmentioned item in the New condition. Thus, there is no reason for concern that the differences we found reflect a difference between looks to an item that was mentioned before and one that was not.

We interpret the remainder of our findings as a manifestation of the RNP in that the Repeated condition led to delayed processing of information relative to the Pronoun condition. This was reflected in the smaller number of fixations to the second mentioned item in the critical sentence, and the increasingly greater number of fixations to the previously mentioned item in the Repeated condition than in the Pronoun condition. We note that, in line with the general claim of the ILH, this effect appeared related to an activated memory representation interfering with processing. However, this interference was associated with a competition driven by the activation of other previously mentioned referents and not, as the ILH had originally claimed, with the activation of broad semantic representations, as participants did not look more often to the Semantic Distractor in the Repeated condition relative to the Pronoun condition. In fact, participants looked *less* often at the Semantic Distractor initially in both conditions than in the New condition. A possible

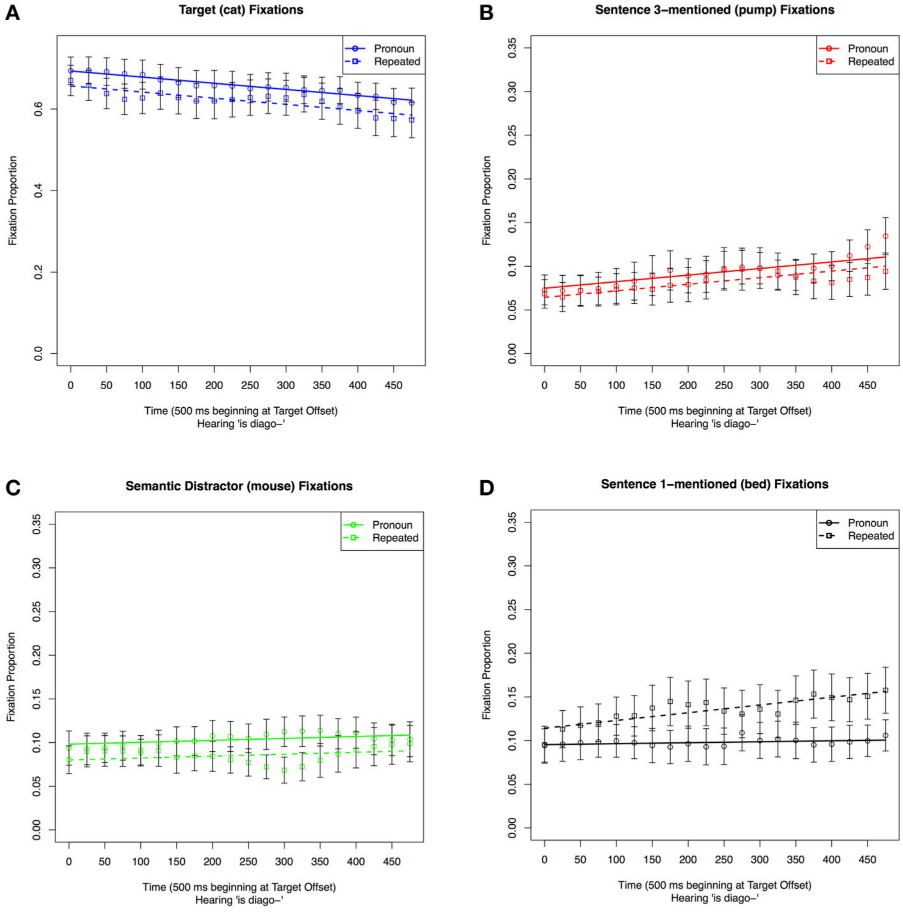


FIGURE 5 | Proportion of fixations to individual pictures on the display in the 500 ms time window starting at the offset of the reference to the Target in Sentence 3. Fixations are graphed by anaphor form condition (Pronoun vs. Repeated): **(A)** Fixations to the pictures of the Target (cat); **(B)** Fixations to the picture of Sentence 3-mentioned (pump); **(C)** Fixations to pictures of the Semantic Distractor (mouse); **(D)** Proportion of fixations to the picture of Sentence 1-mentioned (bed).

objection to this interpretation is that the new and repeated nouns related to different pictures such that when participants heard "cat" for the second time, they had likely looked previously at the picture of the cat and may also looked at the semantically related mouse. As a result, participants may have had less reason to identify and process the pictures of the cat and mouse again, and instead, they looked at the bed. We will return to discuss this objection in the context of the results of Experiment 2.

EXPERIMENT 2

The results of Experiment 1 support the presence of the RNP in spoken language comprehension and the ILH's general claim about the involvement of memory interference related to the activation of other information. These results show that it is the activation of information that was associated with the referent in the previous discourse (Sentence 1-mentioned) that underlies the memory interference in the RNP. One way this finding could be explained is in terms of a cue-based theory of memory retrieval. Specifically, the mention of both referents in Sentence 1 may have created a representation of the two as a cue-retrieval target pair or at least combined some information about the Sentence 1-mentioned item with the discourse representation of the Target. Under a cue-based theory of memory retrieval, this may have resulted in the automatic retrieval of the Sentence 1-mentioned representation upon hearing the Target in Sentence 3, and this irrelevant retrieved information, caused the delay in processing. This retrieval process may have been more effective following repeated references because the extra information in these references may have provided a stronger retrieval cue.

Importantly, in contrast to the specific prediction of the ILH (Almor, 1999), Experiment 1 did not show any evidence of semantic effects in the Repeated condition. Thus, the results of Experiment 1 can be explained on the basis of a general memory mechanism, rather than the activation of pre-existing semantic relations in long-term memory. Given that previous research in reading has shown the involvement of long-term memory semantic relations in the RNP in reading (Almor, 1999; Cowles and Garnham, 2005), we wanted to further explore the absence of a similar effect here. Specifically, we wanted to ascertain whether a pre-existing semantic relation can modulate and interact with the retrieval and activation of the discourse representation of the referent in the future, perhaps causing interference when deciding upon a new referent, when it has been previously activated. For example, if *cat* and *mouse* are mentioned together, can the fact that the two have a pre-existing semantic relationship affect recall of *mouse* when *cat* is mentioned again later, vs. if *cat* and *bed* were originally mentioned together (as in Experiment 1). Experiment 2 therefore tested whether the processing of repeated reference is affected by the strength of the semantic relation between the referent and the information associated with it earlier in the discourse.

The design of Experiment 2 followed Experiment 1 closely except that it manipulated whether the other object mentioned with the Target in Sentence 1 was an unrelated object (the *Unrelated* condition), or the Semantic Distractor, (the *Related* condition). For clarity purposes, we will from now on refer to the unrelated object as *Sentence 1-unrelated*. Sentence 3 in Experiment 2 appeared only in two conditions: *Pronoun* and *Repeated*. The New condition from Experiment 1 was not of interest for the question at hand and therefore was not included. **Table 1B** shows a sample item. Experiment 2 used the same pictorial displays as in Experiment 1.

This experiment aimed to test the following specific predictions:

1. According to a simple cue-based retrieval explanation of reference processing, pre-existing semantic relations should facilitate the retrieval of the representation of the referent. Therefore, there should be more fixations to an item that is semantically related to the target referent (Semantic Distractor, *mouse*) than a comparable unrelated item (Sentence 1-unrelated, *bed*) when it was previously mentioned with the Target. This effect should be stronger for repeated names than for pronouns as repeated names provide a stronger retrieval cue.
2. According to the ILH, the Related and Unrelated Sentence 1 conditions should have a different effect on the processing of potential referents in the Repeated and Pronoun Sentence 3 conditions. Specifically, the interference in processing repeated anaphors is expected to be greater in the Related than in the Unrelated conditions, due to the increased activation of a related previously mentioned item (the Semantic Distractor in the Related conditions) than an unrelated previously mentioned item (the Sentence 1-unrelated in the Unrelated conditions). In the Repeated conditions, this should be reflected in more fixations that increase at a higher rate to the Semantic Distractor in the Related than in the Unrelated conditions. The ILH predicts that in the Pronoun conditions there will be no such differences.

As far as we can tell, DPT does not make any prediction about differences in processing between the Related and Unrelated conditions within either the Repeated condition or the Pronoun condition. Evidence of such differences is therefore unexpected according to DPT, but not necessarily incompatible with it.

Once again, our analyses involved GCAs. All data was preprocessed following the steps outlined in Experiment 1.

METHOD

Participants

Fifty-eight undergraduate students recruited from the University of South Carolina Psychology Department's participant pool participated in this experiment for course credit. All participants provided informed consent in accordance with the University's IRB. All participants were native speakers of American English.

Materials

The pictorial displays used in Experiment 2 were identical to those used in Experiment 1 (see **Figure 1**). To help distinguish between the objects in this experiment, in which a different object was mentioned with the Target in Sentence 1, we refer to the picture labeled *Sentence 1-mentioned* in Experiment 1 as *Sentence 1-unrelated*. The 3-sentence discourses used for Experiment 2 were constructed by altering the experimental items used in Experiment 1. As in Experiment 1, the discourses described the location of the items and always left two items unmentioned until the final referent of Sentence 3 was identified. The first sentence of each discourse appeared in two conditions. Either an Unrelated condition (∼2454 ms), in which, like in Experiment 1, the two referents were unrelated, or a Related condition (∼2603 ms), in which the two referents were related. Indeed, the Unrelated condition simply used the first sentences from Experiment 1. In the Related condition, the second referent mentioned after the Target (*cat*) was the Semantic Distractor (*mouse*) instead of Sentence 1-unrelated (*bed*), and specified its location in relation to the Target. Sentence 2 was identical to Sentence 2 in Experiment 1. Sentence 3 contained only the Pronoun (∼1957 ms) and Repeated conditions (∼2563 ms), introducing a new referent (Sentence 3-mentioned, *pump*) as the second reference.

Verbal stimuli were recorded by the same native female speaker of American English (S.A.P.) and edited using sound editing software. All experimental items included the same version of Sentence 2. Sentences 1 and 3 were recorded separately for each condition. Items were presented in a random order, which differed by participant. Each participant heard each experimental item once such that they responded to six items in each condition. Across all participants, each item appeared in each condition a similar number of times. Experimental items were always true, and 12 of the 48 fillers were also true, such that overall the verbal descriptions in exactly half of the trials were true.

Apparatus

The apparatus used in Experiment 2 was identical to that used in Experiment 1.

Procedure

The procedure and task for Experiment 2 were identical to that of Experiment 1. Response accuracy for the task was again recorded; no participants were removed from analyses due to low accuracy within the task.

TABLE 5 | **Coefficient estimates in the best-fitting quadratic model.**

Coefficient	*Est.*	*Std. Error*	*t*	*p <*
Intercept	0.1794	0.0080	22.341	0.001
Time	−0.1313	0.0207	−6.332	0.001
Time^2	0.0432	0.0153	−2.825	0.01
S1-U	−0.0397	0.0048	−8.225	0.001
S3-M	−0.0851	0.0048	−17.615	0.001
Time*S1-U	−0.0910	0.0216	−4.210	0.001
Time*S3-M	0.0377	0.0216	1.424	*n.s.*
Time^2*S1-U	0.0727	0.0216	3.366	0.001
Time^2*S3-M	0.0738	0.0216	3.414	0.001

Comparisons for proportion of fixations to the Semantic Distractor, Sentence 1-unrelated (S1-U), and Sentence 3-mentioned (S3-M) in Sentence 1 in a 500 ms time window starting 100 ms before the offset of the Target in Sentence 1. Proportion of fixations to the Semantic Distractor provided the baseline group.

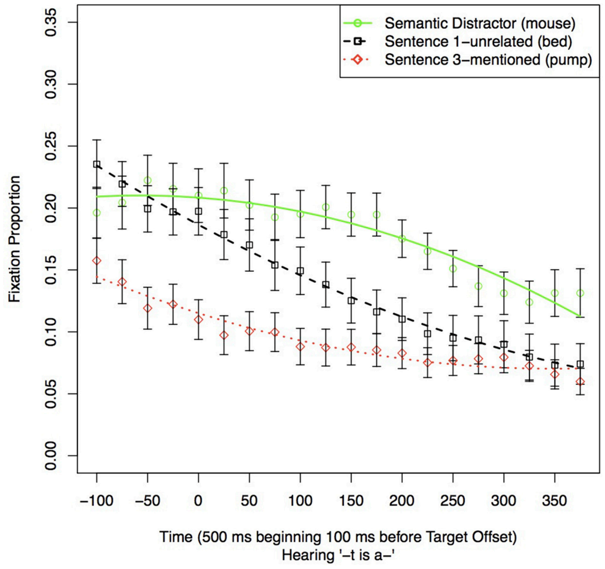


FIGURE 6 | Proportion of fixations to other items in the display in a 500 ms time window starting 100 ms before the offset of the Target reference in Sentence 1. This figure replicates the semantic competitor effect observed in Experiment 1.

RESULTS

Raw eye position data transformation and condition matching were the same as in Experiment 1. Ten participants were removed before the analysis, due to equipment failure or poor calibration during the experiment, leaving 48 participants.

Sentence 1

We tested for a replication of the semantic competitor effect from Experiment 1. We examined fixations during Sentence 1 in the Unrelated condition in the same time window as in Experiment 1. We only included the Unrelated condition because immediately at the offset of the Target, participants already began hearing the location of the second mentioned item, which in the Related condition was the Semantic Distractor. This made it impossible to gauge the effect of semantic relatedness in the Related condition. The results of the analyses are shown in **Table 5** and **Figure 6**. The best-fitting model included effects of condition on both the linear and quadratic time terms. Coefficient estimates were close to those obtained in Experiment 1, with the exception that the effect of the Sentence 3-mentioned on the linear component of Time, which reversed in sign. The quadratic components indicate that the changes in proportion of fixation is different for fixations to the items. Combined, these data indicate a semantic competitor effect similar to the one observed in Experiment 1. The semantic competitor (Semantic Distractor) received more fixations than the other two objects not yet mentioned. The slight difference in results is not unexpected as the Semantic Distractor was never mentioned in Experiment 1, yet here, although not in the trials used to test for the effect, it was mentioned.

TABLE 6 | **Coefficient estimates in the best-fitting model for proportions of fixations to items previously mentioned with the Target in Sentence 1 during Sentence 3 in the 500 ms time window following Target offset in the Repeated condition.**

Coefficient	*Est.*	*Std. Error*	*t*	*p <*
Intercept	0.0858	0.0117	7.308	0.001
Time	−0.0035	0.0196	−0.189	*n.s.*
Unrelated-S1-U	−0.0246	0.0047	−5.184	0.001

Models in the Pronoun condition demonstrated no significant difference. Proportion of fixations at the offset of the Target in the Related condition to the Semantic Distractor provided the baseline group.

Sentence 3

Prediction 1

To test the cue-based retrieval explanation, we carried out analyses comparing fixations to an item previously mentioned with the Target when it was related to the Target (Semantic Distractor, *mouse*) to when it was not (Sentence 1-unrelated, *bed*). We did this separately for the Repeated and Pronoun conditions.

The results of the analyses are shown in **Table 6** and **Figure 7**. For the Repeated conditions, when mentioned together with the Target in Sentence 1, there were more looks to an item that was semantically related to the Target (Semantic Distractor, *mouse*), than to an item that was not (Sentence 1-unrelated, *bed*) (**Table 6**, **Figure 7A**). This shows that a pre-existing semantic relation can modulate the interference caused by items mentioned earlier in the discourse. For the Pronoun conditions (**Figure 7B**) there were no differences in looks to the items dependent on Sentence 1 condition, so while the graph is included for illustrative purposes, a corresponding model does not appear. Thus, there is no evidence for interference caused by a pre-existing semantic relationship on pronoun resolution.

Prediction 2

We carried out analyses comparing fixations to each display item in the Related and Unrelated conditions of Sentence 1,

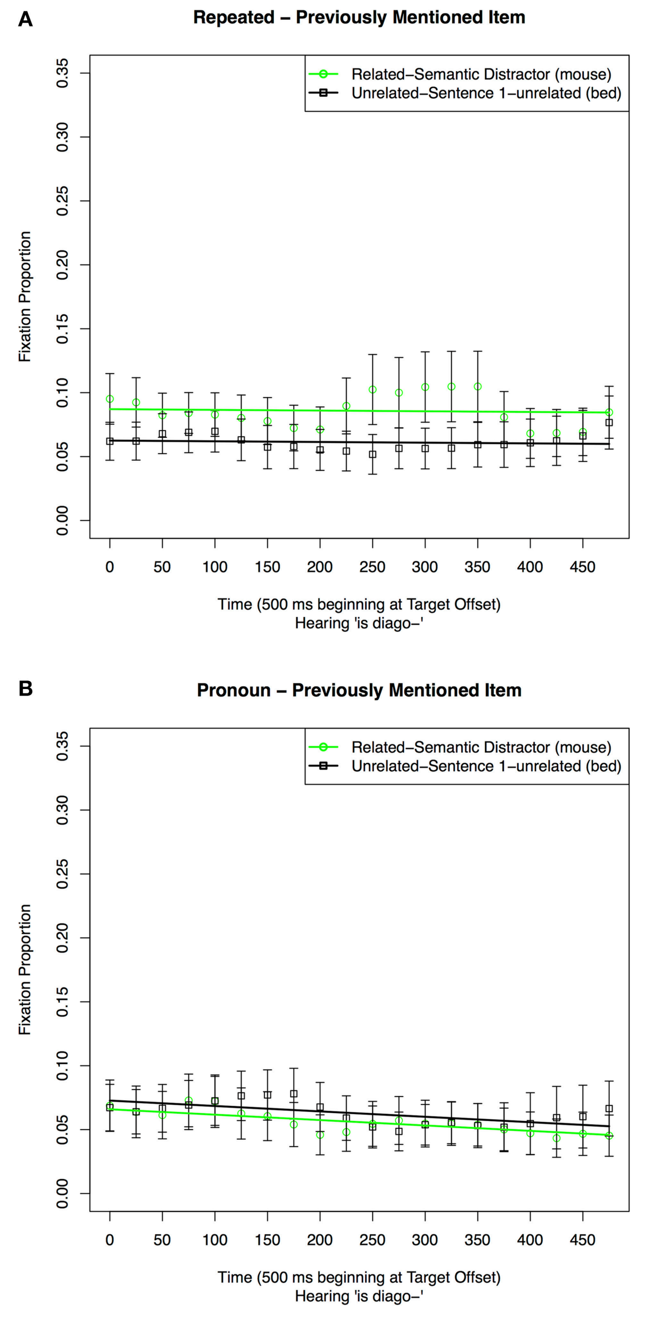


FIGURE 7 | Proportion of fixations to the items in the display previously mentioned with the Target in a 500 ms time window starting at the offset of the Target reference in Sentence 3 in the (A) Repeated and (B) Pronoun conditions.

TABLE 7 | Coefficient estimates for the best-fitting models in the 500 ms time window following Target offset in Sentence 3 of a Repeated (Model 1) or Pronoun (Model 2) reference when Sentence 1 was in the Unrelated vs. Related condition, and proportion of fixations to the (A) Target, (B) Sentence 3-mentioned, (C) Semantic Distractor, and (D, Repeated only) Sentence 1-unrelated served as the outcome.

Coefficient	*Est.*	*Std. Error*	*t*	*p <*
Model 1. Repeated				
(A) TARGET				
Intercept	0.5993	0.0363	16.528	0.001
Time	−0.1438	0.0428	−3.360	0.01
Time^2	−0.0685	0.0237	−2.897	0.01
Unrelated	−0.0328	0.0075	−4.391	0.001
Time*Unrelated	0.1006	0.0335	3.007	0.01
Time^2*Unrelated	0.0533	0.0335	1.593	*n.s.*
(B) SENTENCE 3-MENTIONED				
Intercept	0.1280	0.0172	7.464	0.001
Time	0.0749	0.0351	2.134	0.05
Time^2	0.0260	0.0186	1.395	*n.s.*
Unrelated	−0.0112	0.0059	−1.911	0.06
Time*Unrelated	−0.0054	0.0263	−0.207	*n.s.*
Time^2*Unrelated	−0.0893	0.0263	−3.395	0.001
(C) SEMANTIC DISTRACTOR				
Intercept	0.0858	0.0171	5.022	0.001
Time	−0.0070	0.0344	−0.203	*n.s.*
Unrelated	−0.0029	0.0051	−0.568	*n.s.*
Time*Unrelated	−0.0778	0.0226	−3.438	0.001
(D) SENTENCE 1-UNRELATED				
Intercept	0.0423	0.0096	4.406	0.001
Time	0.0164	0.0127	1.291	*n.s.*
Unrelated	0.0189	0.0042	4.548	0.001
Model 2. Pronoun				
(A) TARGET				
Intercept	0.6135	0.0401	18.060	0.001
Time	−0.0119	0.0489	−0.314	*n.s.*
Unrelated	0.0182	0.0007	−6.238	0.001
(B) SENTENCE 3-MENTIONED				
Intercept	0.1262	0.0154	8.211	0.001
Time	0.0602	0.0322	1.870	0.07
Time^2	0.0212	0.0179	1.181	*n.s.*
Unrelated	−0.0424	0.0057	−7.487	0.001
Time*Unrelated	−0.0700	0.0253	−2.762	0.01
Time^2*Unrelated	−0.0177	0.0253	0.698	*n.s.*
(C) SEMANTIC DISTRACTOR				
Intercept	0.0559	0.0118	4.743	0.001
Time	−0.0322	0.0190	−1.693	*n.s.*
Unrelated	0.0216	0.0043	5.000	0.001
Time*Unrelated	0.0535	0.0193	2.764	0.01

The Related condition served as the baseline.

first in the Repeated conditions and then in the Pronoun conditions. Because our focus in this experiment was on the effect of semantic relatedness, we chose to conduct a separate set of analyses for each type of referential expression. This approach was not used in Experiment 1, in which the informative comparisons were between different reference types.

Repeated condition: fixations to individual objects in the related vs. unrelated condition

Target (cat)

The results of the analyses are shown in **Table 7** Model 1A and **Figure 8A**. The chosen quadratic model included significant effects of condition on the intercept, slope and quadratic Time terms. These reflect participants initially fixating more on the

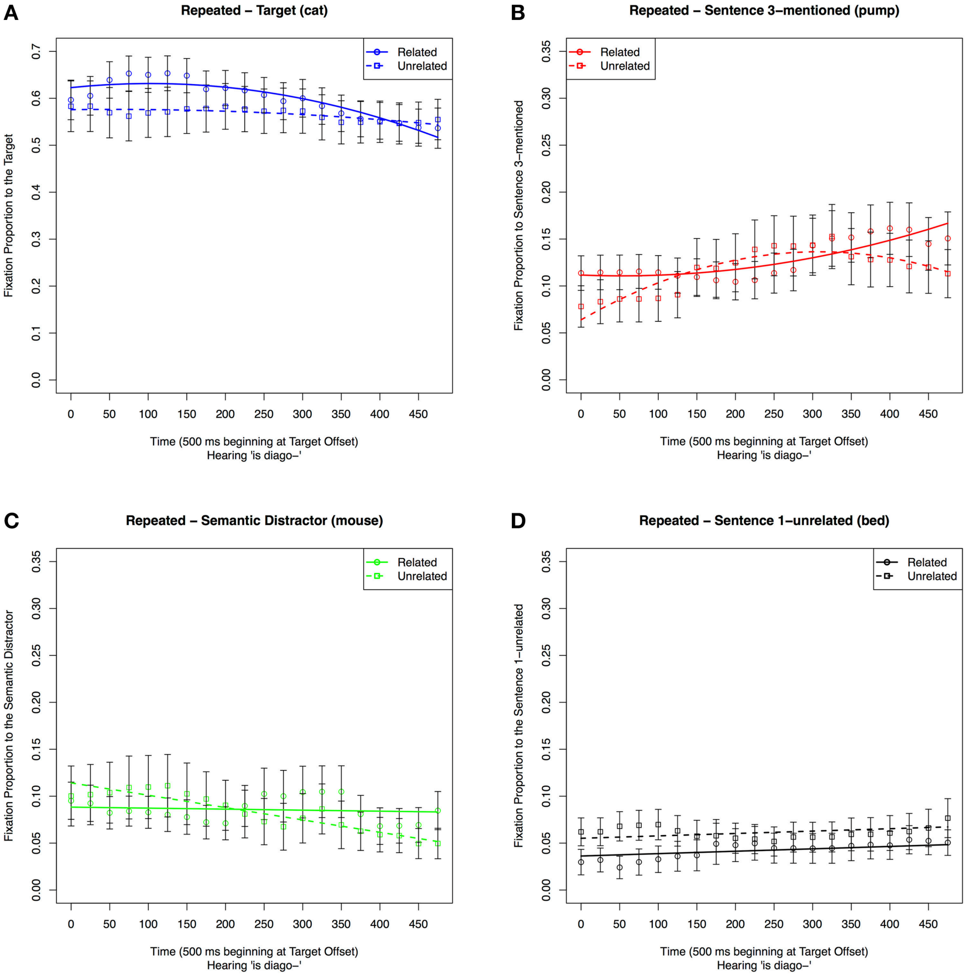


FIGURE 8 | Proportion of fixations in the 500 ms time window following Target offset to the (A) Target, (B) Sentence 3-mentioned, (C) Semantic Distractor, and (D) Sentence 1-unrelated in the Repeated conditions when the Sentence 1 condition was Unrelated or Related.

Target and later fixating away from it sooner in the Related condition than in the Unrelated condition.

Sentence 3-mentioned (pump)

The results of the analyses are shown in **Table 7** Model 1B and **Figure 8B**. Overall there were marginally more fixations to the Sentence 3-mentioned item in the Related condition. However, as shown in the graph and indicated by the quadratic effects of condition on Time, fixations in the Related condition rose over time, while fixations in the Unrelated condition rose and then fell in the same window.

Semantic distractor (mouse)

The results of the analyses are shown in **Table 7** Model 1C and **Figure 8C**. The best-fitting model included a condition effect on the intercept and the slope Time term. While fixations to the Semantic Distractor increased with time in the Related condition, they decreased in the Unrelated condition.

Sentence 1-unrelated (bed)

The results of the analyses are shown in **Table 7** Model 1D and **Figure 8D**. The best-fitting intercept only GCA model for these data included an effect of condition. There were more fixations to Sentence 1-unrelated in the Unrelated condition than in the Related condition.

Pronoun condition: fixations to individual objects in the related vs. unrelated conditions

Target (cat)

The results of the analyses are shown in **Table 7** Model 2A and **Figure 9A**. The selected model included an effect of condition only on the intercept. Thus, as is shown in both table and figure, participants looked more often at the Target in the Unrelated condition than in the Related condition at the Target offset, but there were no differences in the time course of processing. This differs from the Repeated condition, where participants initially fixate on the Target in the Related condition,

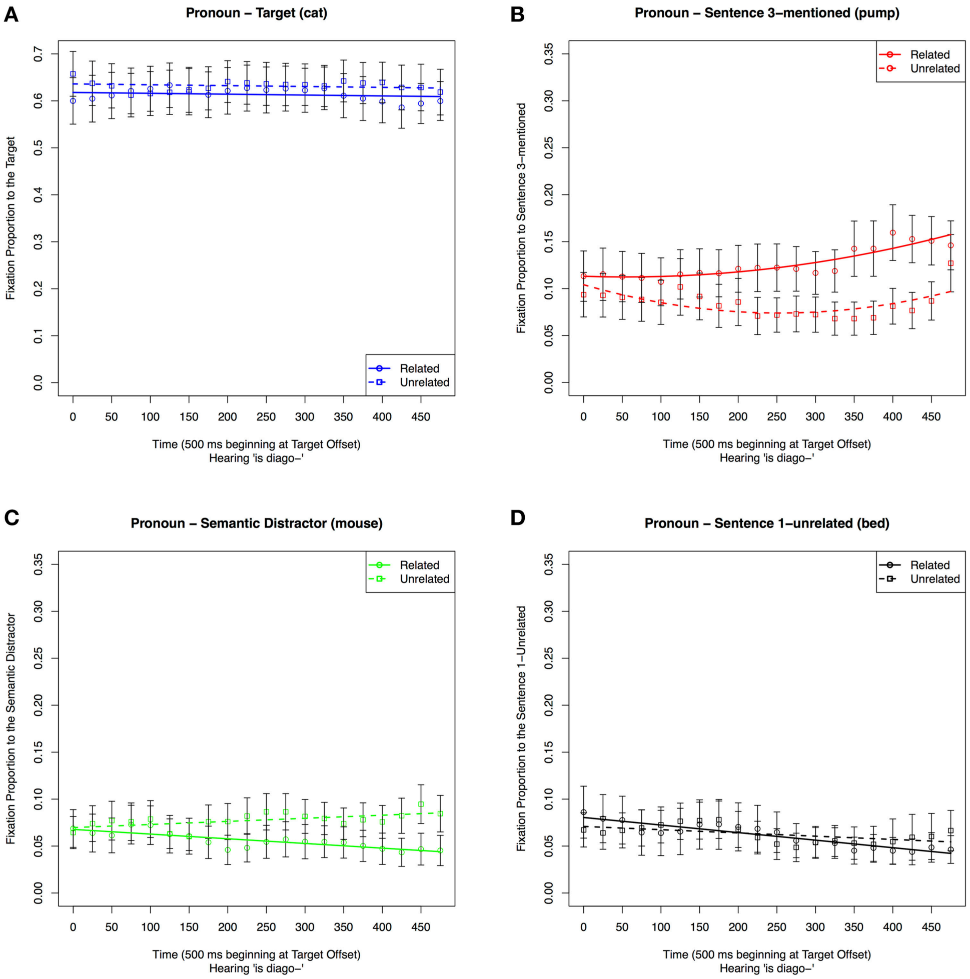


FIGURE 9 | Proportion of fixations to the (A) Target, (B) Sentence 3-mentioned, (C) Semantic Distractor, and (D) Sentence 1-unrelated in the 500 ms time window following Target offset in the Pronoun conditions when the Sentence 1 condition was Unrelated or Related.

then fixate away from it at a quicker rate than the Unrelated condition.

Sentence 3-mentioned (pump)

The results of the analyses are shown in **Table 7** Model 2B and **Figure 9B**. The selected quadratic model included an effect of condition on the slope Time coefficient. As is shown in both table and figure, participants looked more often and at an increased rate at Sentence 3-mentioned in the Related condition than in the Unrelated condition. This differed from the Repeated condition in which the Unrelated condition had an increasing and then decreasing pattern of fixations within the same time window.

Semantic distractor (mouse)

The results of the analyses are shown in **Table 7** Model 2C and **Figure 9C**. The best-fitting model only included a significant effect of condition on the slope of the time parameter, with a quicker rate of fixating away from the Semantic Distractor in the Unrelated than in the Related condition. This differed from the Repeated condition, in which fixations in the Unrelated condition were lower and decreased over time.

Sentence 1-unrelated (bed)

The graphical result of the analysis is shown in **Figure 9D**. While the intercept model was graphed for full comparison purposes as it was the best fit, the model was not significant and is not included. Thus, there were not any differences between looks to Sentence 1-unrelated in the Related vs. Unrelated conditions following pronouns. This differs from the Repeated condition in which the item received more fixations in the Unrelated condition.

Overall these results show that when an item is initially mentioned with the Target, there are differences due to the semantic relation between that item and the Target on the later processing of a reference to the Target. In the case of a repeated reference, previously related items (*mouse*) receive more fixations than unrelated ones (*bed*). Also, for repeated references, mentioning an item related to the Target initially hinders performance and then has a facilitative effect. This could reflect the related item being considered as the next possible referent. In contrast, for pronouns, mentioning an item related to the Target facilitates resolution in comparison to mentioning an unrelated item. This could reflect the quicker dismissal of a related item than an unrelated item as a candidate for being the next possible referent.

DISCUSSION

The results of this experiment show clearly that pre-existing semantic relations between a referent and a previously mentioned item generally facilitate reference resolution in our paradigm. Following both pronouns and repeated definite references, a semantic relation between the target referent and a previously mentioned item facilitated processing. This was reflected in the higher rate of fixations to the referent mentioned next in Sentence 3 (Sentence 3-mentioned; *pump*) in the Related than in the Unrelated conditions at the end of the time window. However, despite the similarity in the effect of semantic relatedness at the end of the time window following pronouns and repeated definite references, there were important differences in the time course of this effect for the two reference types. While semantic relatedness consistently facilitated processing across the entire time window following pronouns, its effect on processing varied following repeated definite references. In particular, following repeated references, the higher fixation rate to Sentence 3-mentioned (*pump*) in the Related compared to the Unrelated condition occurred only in the last part of the time window.

These results support the predictions of the cue-based retrieval view (*Prediction 1*) in that following repeated names, but not pronouns, there were more fixations to the Semantic Distractor than to Sentence 1-unrelated when each was mentioned with the Target in Sentence 1 (**Figure 7**). The results are compatible with the ILH (*Prediction 2*) in that, following repeated names, there were more fixations that decreased at a slower rate to the Semantic Distractor in the Related than in the Unrelated condition. Also in line with this prediction, this pattern reversed following pronouns in that there were fewer fixations that decreased at a higher rate to the Semantic Distractor in the Related than in the Unrelated condition. It thus appears that for pronouns, semantic relatedness of a previously mentioned item resulted in the quicker rejection of inappropriate referents. For repeated names, the process was a bit more complex. When a previously mentioned item was semantically related to the referent, it was briefly considered a possible referent of the repeated reference, but was quickly discarded.

An alternative explanation for why participants often looked at the Semantic Distractor when they heard the Target may be due to automatic spreading activation between related concepts/words. In other words, participants may have suppressed "the mouse" as a potential antecedent for the "cat," but may have nevertheless looked at the picture of "mouse" regardless of whether it could be a potential antecedent. Because pronouns are semantically related to neither the Target nor the Semantic Distractor, this did not happen after pronouns. While this interpretation provides a possible explanation for the results of Experiment 2, it is incompatible with the results of Experiment 1 in which the semantic distractor received more fixations in the Pronoun than in the Repeated condition. The results of Experiment 1 thus indicate that the effect in Experiment 2 is clearly related to the fact that the Semantic Distractor was mentioned in the discourse.

It should be noted that the activation of the previously unmentioned referents during the processing of repeated references is also compatible with the main tenant of DPT, which is that repeated reference is initially interpreted as a new reference. However, the finding that these activations are sensitive to the semantic relations between previously mentioned referents, is not predicted by DPT.

In the discussion of Experiment 1, we described an alternative explanation for the increased looks to the previously unmentioned item in the Repeated condition relative to the New condition. According to this alternative explanation, this difference merely reflected the greater likelihood that the Target and the Semantic Distractor were already looked at in comparison to the unmentioned referent. This explanation is incompatible with the finding in the current experiment that semantic relatedness increased this effect rather than weakened it as this alternative explanation would predict (given that participants were more likely to have previously looked at the Semantic Distractor than at Sentence-1-Unrelated).

GENERAL DISCUSSION

Overall, our results indicate that an effect similar to the RNP observed in self-paced reading also occurs in spoken language comprehension. Our results also allow us to understand the time course and possible memory basis of this effect better than in previous reading studies. In the current study, this effect was reflected in delayed fixations to the second referent mentioned in the critical sentence following a repeated reference relative to a pronoun. Use of the VWP in conjunction with GCA techniques allowed us to examine the fine time course of the underlying processes, and demonstrate that the RNP is associated with discourse integration, which is delayed beyond the initial processing of the reference. Our results further show that such delays are related to the memory activation of discourse representations, and that this activation is influenced by a combination of previous mentions, semantic relations, and reference type. To our knowledge, our study is the first to use GCA analyses to better understand the time course of discourse reference in spoken language comprehension. We believe we have shown that using this type of analysis can be profitable for the understanding of these processes.

Our results provide mixed evidence regarding DPT (Gordon and Hendrick, 1998). In contrast to DPT's core claim that repeated references are processed like new references, Experiment 1 revealed that the two kinds of reference are processed differently. In that experiment, repeated references increased fixations to previously mentioned items, but new references increased fixations to items that were not previously mentioned. Nevertheless, the results of Experiment 2 provided some support for DPT in finding that previously unmentioned items were considered possible referents for a repeated reference. However, DPT does not predict the finding that this consideration was influenced by the semantic relation between the unmentioned items and the target referent. While this finding is not plainly incompatible with DPT, it does place this theory at a disadvantage relative to theories that do specifically predict semantic effects. Overall, while DPT's claim that repeated and new references are processed alike may be too simplistic, a weaker version of this claim may be true. The processing of repeated references may generally involve the consideration of previously unmentioned items, but mentioned items are considered first, and semantic representations play a role.

Our results support the general claim of the ILH (Almor, 1999, 2000, 2004; Almor and Nair, 2007) that the RNP is related to memory interference that delays the integrative processing of the reference. At the same time, the results also help clarify the nature of this memory interference. Specifically, our results show that this interference reflects the activation of prior information associated with the referent at the expense of ongoing discourse integration. Experiment 2 further showed that semantic relations play a role in this interference. When the two items that were mentioned together in Sentence 1 were semantically related, a pronoun reference was processed quicker and a repeated reference was processed slower. This suggests that processing both pronouns and repeated references involves activation of semantic discourse representations, although this activation affects the two reference types differently.

These findings can be explained in a cue-based memory framework. When two items are mentioned together, their discourse representations are more strongly connected when they are semantically related than when they are not. Therefore, a later mention of one of the items causes a quicker and stronger activation of the other when the two are related. This appears to have a different effect on the processing of pronouns and repeated names. Although it is possible that this is related to the consideration of the reasons for why, in the repeated condition, a repeated name has been used rather than a pronoun, this does not explain the specific patterns of results or provide any additional information about the underlying memory mechanism. Instead, we hypothesize that processing pronouns involves picking the most salient referent while actively suppressing other possible referents. The quicker activation of the representation of the other item in the related case allows for its quicker suppression as well, relative to the unrelated case. In contrast, processing repeated references involves a competition between the activated possible referents. Therefore, the stronger activation of a mentioned item when it is related to the Target relative to when it is not, leads to greater competition, causing a delay in processing. This explanation is compatible with the general claim of the ILH that the RNP reflects memory interference between semantic representations. However, unlike in previous work on the ILH, the interference here is caused by considering alternative and upcoming referents rather than by direct memory interference between the representations of the referent and the current reference.

The difference between the interference found in this study and the interference claimed by the ILH could be attributed to several factors. The first is the type of manipulation used in the present study vs. previous studies of semantic effects on reference processing. In contrast to the present study, several previous studies manipulated the semantic distance between a referential expression and the original mention of the referent (e.g., Sanford and Garrod, 1981; Garnham et al., 1997; Almor, 1999; van Gompel et al., 2004; Cowles and Garnham, 2005). Moreover, these studies focused on a hierarchical semantic overlap between the reference and the previous mention (e.g., *robin-bird* or *bird-animal*), whereas the semantic relations we examined here were based on a broader notion of semantic relatedness that did not involve hierarchical relations (e.g., *hammer-nail*). Thus, the interference found in the present study does not preclude the existence of other forms of interference, such as between semantically overlapping representations of referents and references.

In addition to the importance of these results for the two theories we tested here, we believe that our findings about the memory processes and activations associated with the different types of reference are novel and provide a meaningful empirical contribution to the literature. Overall, we have shown that reference processing reflects underlying memory representations and processes that, in line with general theories of memory, are affected by semantic relations and previous mention. A closer semantic relation between a previously mentioned item and a co-mentioned referent results in a stronger activation of the co-mentioned referent when a subsequent reference is encountered. For pronominal references, this stronger activation allows quicker suppression of the co-mentioned referent and therefore a quicker identification of the correct referent. In the case of repeated references, this stronger activation results in increased competition, which interferes with identifying the correct referent. Therefore, although pronouns and repeated references are processed differently, these differences can still be captured by general memory principles such as interference, suppression and competition. Finally, we believe that our novel use of GCA to study the processing time course of referential expressions provides a methodological contribution to the literature.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct, and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the University of South Carolina Research and Productive Scholarship Fund, NSF award BCS 0822617, and NIH R21AG030445. We thank Eileen Yee for sharing her stimuli. We thank Dan Mirman for sharing his R code and analysis steps online. We also thank Mary Katherine Allawos, Eric Bowers, Lauren Hodge, and Jeremy May for help with testing and Veena Nair for comments and discussion. Parts of this research were presented at the 2007 Annual CUNY Conference on Human Sentence Processing, San Diego, CA, USA, and the 2007 annual North Carolina Cognition Conference, Chapel Hill, NC, USA.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.00214

REFERENCES

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition: evidence for continuous mapping models. *J. Mem. Lang.* 38, 419–439. doi: 10.1006/jmla.1997.2558

Almor, A. (1999). Noun-phrase anaphora and focus: the informational load hypothesis. *Psychol. Rev.* 106, 748–765. doi: 10.1037/0033-295X.106.4.748

Almor, A. (2000). "Constraints and mechanisms in theories of anaphor processing," in *Architectures and Mechanisms for Language Processing*, eds M. W. Crocker, M. Pickering, and C. Clifton, Jr. (New York, NY: Cambridge University Press), 341–354.

Almor, A. (2004). "A Computational Investigation of reference in production and comprehension," in *Approaches to Studying World-situated Language Use: Bridging the Language-as-product and Language-as-action Traditions*, eds J. C. Trueswell and M. K. Tanenhaus (Cambridge, MA: MIT Press), 285–301.

Almor, A., and Eimas, P. D. (2008). Focus and noun phrase anaphors in spoken language comprehension. *Lang. Cogn. Process.* 23, 201–225. doi: 10.1080/01690960701330936

Almor, A., and Nair, V. A. (2007). The form of referential expressions in discourse. *Lang. Linguist. Compass*. 1, 84–99. doi: 10.1111/j.1749-818X.2007.00009.x

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London; New York, NY: Routledge.

Arnold, J. A., Tanenhaus, M. K., Altmann, R. J., and Fagnano, M. (2004). The old and, theee, uh, new: disfluency and reference resolution. *Psychol. Sci.* 9, 578–582. doi: 10.1111/j.0956-7976.2004.00723.x

Baddeley, A. (1992). Working memory. *Science* 255, 556–559. doi: 10.1126/science.1736359

Baddeley, A. (1996). The fractionation of working memory. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13468–13472. doi: 10.1073/pnas.93.24.13468

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-effects Models using Eigen and S4. R Package Version 1.1-7.* Available online at: http://CRAN.R-project.org/package=lme4

Boiteau, T. W., Malone, P. S., Peters, S. A., and Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *J. Exp. Psychol.* 143, 295–311. doi: 10.1037/a0031858

Chen, Q., and Mirman, D. (2015). Interaction between phonological and semantic representations: time matters. *Cogn. Sci.* 39, 538–558. doi: 10.1111/cogs.12156

Cowles, H. W., Garnham, A., and Simner, J. (2010). Conceptual similarity effects on working memory in sentence contexts: testing a theory of anaphora. *Q. J. Exp. Psychol.* 63, 1218–1232. doi: 10.1080/17470210903359198

Cowles, W. H., and Garnham, A. (2005). Antecedent focus and conceptual distance effects in category noun-phrase anaphora. *Lang. Cogn. Process.* 20, 725–750. doi: 10.1080/01690960400024624

Dahan, D., Tanenhaus, M. K., and Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *J. Mem. Lang.* 47, 292–314. doi: 10.1016/S0749-596X(02)00001-3

Garnham, A., Oakhill, J., and Cain, A. K. (1997). The interpretation of anaphoric noun phrases time course, and effects of overspecificity. *Q. J. Exp. Psychol. A* 50, 149–162. doi: 10.1080/713755687

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cogn. Sci.* 17, 311–348. doi: 10.1207/s15516709cog1703_1

Gordon, P. C., and Hendrick, R. (1998). The representation and processing of coreference in discourse. *Cogn. Sci.* 22, 389–424. doi: 10.1207/s15516709cog2204_1

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: a framework for modeling the local coherence of discourse. *Comput. Linguist.* 21, 203–226.

Gundel, J., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Lang. Cogn. Process.* 69, 274–307. doi: 10.2307/416535

Henderson, J. M., and Ferreira, F. (2004). "Scene perception for psycholinguists," in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World,* eds J. M. Henderson and F. Ferreira (New York, NY: Psychology Press), 1–58.

Kuznetsova, A., Brockhoff, B., and Christensen, R. H. B. (2014). *lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (lmer Objects of lme4 Package). R Package Version 2.0-11*. Available online at: http://CRAN.R-project.org/package=lmerTest

Ledoux, K., Gordon, P. C., Camblin, C., and Swaab, T. (2007). Coreference and lexical repetition: mechanisms of discourse integration. *Mem. Cogn.* 35, 801–815. doi: 10.3758/BF03193316

Long, J. D. (2012). *Longitudinal Data Analysis for the Behavioral Sciences Using R*. Los Angeles, CA: Sage.

Mirman, D. (2014). *Growth Curve Analysis and Visualization Using R*. New York, NY: CRC Press.

Mirman, D., Dixon, J. A., and Magnuson, J. (2008). Statistical and computational models of the visual world paradigm: growth curves and individual differences. *J. Mem. Lang.* 59, 475–494. doi: 10.1016/j.jml.2007.11.006

Peters, S., and Almor, A. (2006). "The repeated-name penalty observed in reference to concrete objects," in *The 19th Annual CUNY Conference on Human Sentence Processing* (New York, NY).

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Sanford, A. J., and Garrod, S. C. (1981). *Understanding Written Language.* Chichester: John Wiley & Sons.

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime Reference Guide.* Pittsburgh, PA: Psychology Software Tools, Inc.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

van Gompel, R. P. G., Liversedge, S. P., and Pearson, J. (2004). "Antecedent typicality effects in the processing of noun phrase anaphors," in *The On-line Study of Sentence Comprehension: Eyetracking, Erps, and Beyond,* eds J. M. Carreiras and C. Clifton (Brighton: Psychology Press), 119–137.

Watkins, O. C., and Watkins, M. J. (1975). Buildup of Proactive Inhibition as a Cue-Overload Effect. *J. Exp. Psychol.* 104, 442–452. doi: 10.1037/0278-7393.1.4.442

Yee, E., and Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *J. Exp. Psychol.* 32, 1–14. doi: 10.1037/0278-7393.32.1.1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on