Ordinary memory processes in the design of referring expressions

Kieran J. O'Shea*, Caitlyn R. Martin, Dale J. Barr

Institute of Neuroscience and Psychology, University of Glasgow, 62 Hillhead St., Glasgow G12 8QB, United Kingdom

Abstract

How do speakers produce referential descriptions that satisfy addressees' informational needs during real-time conversation? A recent proposal is that ordinary memory processes can serve as a proxy for the consideration of common ground. But this is only possible if speakers encode and access sufficiently detailed memory representations. We tested this proposal by having speakers describe referents in contexts varying in perceptual similarity to previous contexts in the dialogue. Based on the analysis of a total of 4,817 descriptions from 112 speakers over three experiments, we found little evidence that contextual similarity modulated the informational content of speakers' descriptions, regardless of whether that similarity was based on configurational cues (Exps. 1 and 2), or on the perceptual experience of interacting with a conversational partner (Exp. 3). In contrast, speakers did modulate their descriptions when their beliefs about the addressee changed, even when the perceptual match between encoding and retrieval contexts was identical. This suggests that the episodic representations accessed during message generation may be too impoverished to serve as an effective proxy for common ground.

*Thanks to David Ralston and Sophie MacAskill for assistance with pilot studies, and to Holly Branigan and Simon Garrod for comments on the PhD thesis on which this article is based. Code and data are available at https://osf.io/89g5b. This work was made possible by a Doctoral Training Fellowship to Kieran J. O'Shea from the UK Economic and Social Research Council.

*Corresponding author.

When planning spoken referential expressions, speakers face many potential choices about what to say and how to say it. What psychological factors govern these choices? It is generally assumed that when planning messages, speakers are beholden to the cooperative principle that underlies conversational interaction (Grice, 1975). Accordingly, speakers should formulate references that convey no more and no less information than their addressee would need to identify the referent of the expression against the background of their common ground: the set of assumptions, beliefs, and knowledge that is shared, and known to be shared with their interlocutors (Clark and Marshall, 1981). Because of the complex nature of reasoning about common ground—i.e., about what you know that I know that you know, etc.—it is assumed that interlocutors assess common ground using a set of simpler *co-presence heuristics*. For instance, interlocutors who can mutually see a vase on a table can assume the vase to be part of their common ground, based on its *physical co-presence*. It is also assumed that interlocutors keep track of things that have been said during a conversation—for instance, what objects have been mentioned and how they were described—and that these things form part of their common ground by virtue of their *linguistic co-presence*. Finally, they can assume certain information to be part of their common ground based on their mutual belonging to a common community (e.g., sports club, internet group, family, etc.) through the heuristic of *community membership* (Clark and Marshall, 1981).

Constructing expressions on the basis of this common ground—a process known as *audience design* (Clark and Murphy, 1982)—can help promote successful communication. However, within the cognitive demands of real-time conversation, common ground can be too uncertain or computationally demanding to estimate and track. It would be expected, then, that speakers use certain shortcuts to make reference generation more efficient.

One such shortcut is to plan utterances relying only on information salient to the self—i.e., without first checking whether it forms part of the common ground—and to allocate cognitive resources to monitoring and adjusting the contextual appropriateness of the expression under construction (Horton and

Keysar, 1996). For instance, a speaker might generate a plan to refer to a candle as *the small candle* to distinguish it from a larger one she knows about, or because that is the way that she referred to this object previously in the discourse, making this description salient. She may rely on the knowledge of the larger circle or the existing precedent simply because they are available during the message generation process (Dell and Brown, 1991), not necessarily because they are part of common ground with the addressee. Using common ground may require considerable time and cognitive resources (Horton and Keysar, 1996; Rossnagel, 2000), which can be in short supply in real-time conversation. This *perspective adjustment* strategy of egocentric planning while using common ground during self-monitoring (Horton and Keysar, 1996) compromises accuracy for efficiency, but may be effective under the assumption that interaction brings interlocutors' perspectives into alignment (Pickering and Garrod, 2004). Moreover, relying on egocentrically available information may not be a huge risk since communicative situations allow for the collaborative detection and resolution of miscommunication (Fussell and Krauss, 1992).

Supporting this view, egocentrically-available information reliably impacts production on a variety of levels, including for example reference generation (Wardlow Lane et al., 2006), syntax (Ferreira and Dell, 2000), and the use of pronouns (Fukumura and van Gompel, 2012). But what exactly is this 'egocentric' information, how does it become accessible in the first place? Unpacking what representations and processes are responsible for making information available in a timely fashion should be an important goal for any theory of language production.

There are likely to be many different kinds of factors that influence the availability of information during reference generation. One appealing possibility is that availability is partially determined by ordinary processes of memory encoding and retrieval. This *ordinary memory view* (Horton and Gerrig, 2005) assumes that everyday memory processes can serve as a proxy for common ground, activating potentially relevant information naturally and effortlessly through cue-driven, parallel search processes often characterized in terms of "resonance" (e.g., Hintzman, 1986).

Consider a speaker who wishes to refer to referent R (e.g., a candle) in context C (e.g., the speaker's living room, while speaking to a friend) and finds it necessary to use expression E (e.g., "the small candle") which includes the modifier 'small' because of features of the context (e.g., the presence of a larger candle). Models of expertise and skill development (Logan, 1988; Logan and Etherton, 1994) suggest that constructing an expression for R will at first be done 'algorithmically'—that is, through deliberative mechanisms involved in everyday problem solving. The cognitive operations and representations involved in this initial assembly become stored in memory as an episodic trace: a *processing episode* (Logan, 1988). With repetition, the process of generating an expression will become increasingly automatized—specifically, the production process will increasingly rely on the wholesale retrieval of expression E from memory on the basis of R and C, which act as retrieval cues. To the extent speakers attend to features of the communicative context during this process (e.g., the environment and audience), these features will become part of the stored processing episode that link the problem situation (i.e., referring to R in context C) to the output solution (the expression E), with each repetition strengthening these links.

Such a view of message generation takes advantage of the key memory principle of *encoding specificity*: stored information becomes available in proportion to the overlap between encoding and retrieval environments (Tulving and Thomson, 1973). The retrieval process may activate various candidate expressions that the speaker has used for the same or similar targets in the past (e.g., "the candle", "the red candle", "the small candle"). Following this principle, the extent to which various candidates are activated will depend on the similarity between the current communicative situation and past situations in which these expressions have been used, providing a kind of automatic route for speakers to produce contextually appropriate references.

Insomuch as similar situations require similar referential expressions, ordinary memory processes may offer a shortcut to successful communication, enacting a process referred to as *attribute substitution* in the judgment and decision-making literature (Kahneman and Frederick, 2002). Attribute substitution is likely to occur whenever a *target attribute* that is needed for a judgment is effortful to compute and there is a *heuristic attribute* available that is correlated with the target attribute, allowing the decision maker to use the latter as a proxy for the former. In the case of reference generation, prior theorizing suggests that the strength of a memory signal associated with a particular referring expression could be a heuristic attribute that can be substituted for the target attribute of common ground. When the speaker attends to a referent with a referential goal, various candidate expressions would become available through memory resonance processes, and the speaker could assess their relative appropriateness in the context through their relative activation strengths (Gann and Barr, 2014). Similarly, Horton and Gerrig (2016) propose that during early stages of utterance generation, retrieval strength could provide a primitive form of *commonality assessment*, with strength of activation providing a surrogate for more explicit computation of common ground.

To the extent that memory associations correlate with common ground, ordinary memory processes could make communicatively relevant information available at minimal cognitive cost. A key prediction of the memory-based view is that conversational partners themselves can act as memory cues, such that the perceptual experiences arising through interactions with a given partner (e.g., the quality of their voice, their appearance, or interaction style) become associated with information that has been shared during those interactions, such that each encounter with the interaction partner re-instantiates shared information.

Although the ordinary memory view has been influential, there is currently little understanding of how, and how much, ordinary memory processes impact information selection during reference generation. A common approach to test the ordinary memory view is to manipulate factors known to affect aspects of memory and assess whether they impact language processing. For instance, Horton and Gerrig (2005) demonstrated that factors affecting memory encoding also impact reference generation, implicating memory as an underlying mechanism. A later study by Ahn and Brown-Schmidt (2020) looked at retrieval rather than encoding, testing the hypothesis that partner effects in referential descriptions should be larger when partner-associated information must be recalled explicitly. Although this hypothesis was not supported, speakers did use longer descriptions of old referents when speaking to new addressees than when they continued speaking about these same referents to old addressees. This finding could be explained in terms of memory processes—specifically, by assuming that new partners cued retrieval of old descriptions less strongly than the old partners—but it could also be explained by assuming speakers just used common ground at some stage during utterance production.

Another approach for investigating ordinary memory processes in language processing is to look at whether re-instantiating previous referential contexts facilitates performance. In picture naming study, speakers were faster to name pictures when speaking to the same partner than when speaking to a different partner (Horton, 2007). This suggests that speakers not only associated previously-produced descriptions with pictures, but also with the prevailing context, including information about the identity of the addressee. However, a later replication attempt called this finding into question (Brown-Schmidt and Horton, 2014). But even assuming the original effect is real, showing enhancement of a performance aspect of production (speech onset time) would fall short of supporting the most important claim of the ordinary memory account for production: that ordinary memory processes can stand in for common ground in the determination of the informational content of an expression.

In short, although there has been some support for the ordinary memory processes in message generation, there is little direct evidence for the key assumption that it can serve as a proxy for common ground. Ordinary memory can only be an effective proxy if two conditions are met: (1) people retain detailed information from previous referring episodes, and (2) these detailed memory representations are accessible to, and taken into account during, the information selection process. The first assumption is well-supported by a large priming literature indicating that people do retain detailed information from past episodes that can influence future behavior, even after long delays (Tulving and Schacter, 1990). However, it is possible that these types of priming processes operate somewhat independently of message planning, influencing aspects of performance (such as speech onset latency) while wielding little or no impact on the selection of information. Thus, it is critical to assess not only the presence of these largely implicit factors, but also to measure their impact on the informational content of speakers' references. These are the main goals of the present set of studies.

Overview of Experiments 1–3

Following work by Brennan and Clark (1996) and Gann and Barr (2014), the logic of the current investigation was to entrain speakers on a referential expression E for referent R in training context C, and then to measure aspects of production in a test context C' that required a different expression, E'. For example, in the training context a speaker might refer to a particular R using expression E, "the small candle", to distinguish it from another larger candle in the referential array. To establish and strengthen memory associations, during a training phase, they would entrain on describing R using E multiple times in context C (or a context similar to it). To avoid direct repetition across trials, these entrainment trials were interleaved with entrainment trials for other referents. Later this same candle would appear in a test context C' where the larger candle was absent. When speakers in C' call the referent "the small candle" (E) instead of simply "the candle" (E') they do so because they are relying on memory instead of the information available in the display. Because we are interested in memory effects, the misspecification rate in C' provides the critical data for our study.

To obtain direct evidence for the ordinary memory view, our study goes beyond previous studies by varying not only the informational requirements from C to C', but also perceptual characteristics that could affect memory retrieval independently of common ground. To the extent ordinary memory processes influence information selection, speakers should be increasingly likely to retrieve

E in the test context C' as a direct function of the perceptual similarity between C' and the training context C, leading in turn to a higher misspecification rate. This key prediction of the ordinary memory view falls out of the encoding specificity principle (Tulving and Thomson, 1973): similarity between encoding and retrieval contexts facilitates retrieval.

To test this prediction, we used the following basic paradigm. Across three experiments, speakers and addressees sat at separate computer screens and engaged in referential communication about shared images. This made it possible to change features of the speaker's display between training and test independently of the addressee's display. In the first two experiments, we varied the similarity in the physical arrangement of images between training and test. In the context of these experiments, the position of a referent on the display was communicatively irrelevant, since speakers believed that listeners viewed a different arrangement of the same objects. In both of these experiments, in addition to recording speakers' descriptions, we also measured their implicit memories for the training displays by tracking eye movements at test. While the implicit measures suggest retention of training display information, there was little evidence for any impact of these implicit memories on speakers' tendency to re-use the descriptions acquired during training.

In the third experiment, speakers spoke to two separate addressees over a video link, which allowed us to manipulate which partner appeared on the screen independently from which partner they were actually speaking to. On certain test trials, speakers viewed a video feed of a partner to whom they had been speaking during training, although they were aware that they were in fact speaking to an addressee who remained off-screen. By independently manipulating who the speaker saw from who they were actually speaking to, we directly tested the idea that partners can serve as a memory cue during message generation. We also independently measured whether speakers kept track of common ground. Although speakers' descriptions were influenced by the identity of the addressee, there was little evidence that the partner they looked at influenced the content of their speech.

The methods and analysis protocols for all three experiments were pre-registered at the Open Science Framework (OSF). The master repository for this project is available at https://osf.io/89g5b, which includes links to pre-registration documents for each experiment, as well as data, code, and a software container providing all necessary infrastructure to reproduce our findings.

Experiment 1

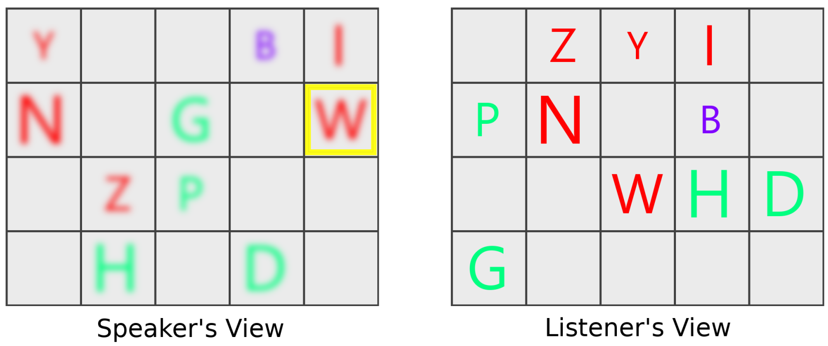

Figure 1: Example test trial displays for Experiment 1 from the Contrast-Singleton condition. The target letter (W) is surrounded by a yellow highlight. The foil letter (Y) appears in the upper left square of the Speaker's view. Note that the Speaker's display was deliberately blurred to make it more difficult to use peripheral vision.

In Experiment 1, speakers referred to capital letters of the alphabet in displays containing multiple distractor letters varying in size and color (Figure 1). Following previous experiments (Brennan and Clark, 1996; Gann and Barr, 2014), we induced referential overspecification using *training* and *test* trials with different informational requirements. For example, during training, a speaker might refer to the letter 'W' as "the large W" to distinguish it from a smaller-sized 'W' in the same array. In the corresponding test trial, the smaller *competitor* W would be replaced by a *foil* letter—a letter with a different identity but that was visually similar to W (e.g., Y) and that also had the same color and size as the competitor (as shown in Figure 1). For all targets in the experiment, size was the only dimension distinguishing the target from the competitor.

We refer to this condition, where the target (e.g., 'W') appeared in training as part of a contrast including a smaller 'W' and as a singleton on the test trial, as the *Contrast-Singleton* condition. To the extent speakers rely on their memory of the training trials, they should overspecify the target at test. Also—and, departing from previous experiments—we included a *Singleton-Contrast* condition intended to induce speakers to *underspecify* referents. Including this condition made the task less repetitive and predictable. In this condition, the target letter appeared as a singleton during training (i.e., with the foil) leading speakers to entrain on the base noun ("the W"), while at test the same target would appear as part of a size contrast with a letter of the same category. If speakers continued using the base description at test, they would provide too little information for the listener to identify the referent. Together, these two conditions formed the levels of a single within-participant factor, *Shift Direction*. Previous research has claimed that speakers are more likely to overspecify than underspecify referents, and even that providing redundant information may sometimes be helpful for listeners Nadig and Sedivy (2002); however, other evidence suggests that unnecessary pre-nominal modifiers may impair listener's comprehension Engelhardt et al. (2011). Although testing these claims was not a main goal of our study, it provides a useful opportunity to clarify some of the issues behind over- versus under-specification.

The critical question in this study was the extent to which the misspecification of target letters varied as a function of the perceptual similarity between the training and test trials. The position of letters in the array was never communicatively relevant, because speakers knew that the listener viewed the same letters, but in a different, unknown arrangement. To test the ordinary memory account, we randomly generated test grids to use as prototypes, which we spatially distorted to create the training trials (Figure 2), as in the classic memory study by Posner and Keele (1968). There were two levels of the within-subject factor of *Distortion*: *Low Distortion*, where the training trials were highly spatially similar to the test prototype; and *High Distortion*, they were less spatially similar to the prototype. To the extent ordinary memory processes influence information selection, speakers in the *Low Distortion* condition should be more

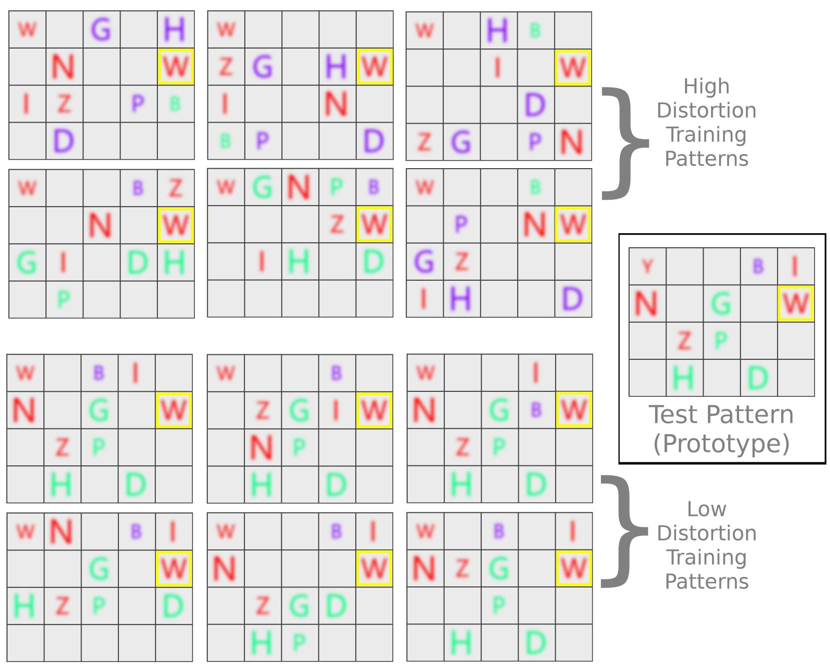

Figure 2: Example prototype with six distortion patterns in each of the two Distortion conditions, Experiment 1. In the High Distortion condition, there was relatively more variability in the letter position (and color) across training patterns; for instance, note that the letter D moves around more and changes color more often.

likely to misspecify referents, because the association between the test display, the referent, and the expression used at training should be stronger and thus more likely to be retrieved.

We chose spatial arrangement as a memory cue because it is widely known that people retain episodic traces of spatial configurations, and that such traces influence aspects of processing; for instance, people are faster to locate a search target when the display configuration is similar to a previous trial (Chun and Jiang, 1998). Similarly, when viewing familiar scenes, people scan them differently, sampling familiar elements less frequently (Ryan et al., 2000). We also used it as a cue because we could easily measure memory for spatial configuration using eyetracking. Because the test display is the prototype of the distortions viewed during training, it should seem more familiar to speakers in the *Low Distortion* condition, and we should expect different scanning patterns relative to the *High Distortion* condition. To the extent we find these differences without changes in speakers' misspecification rate, this would suggest that implicit memory may operate somewhat independently from message planning.

Our main pre-registered prediction was that speakers would be more likely to misspecify referents in the *Low Distortion* than in the *High Distortion* condition. We also predicted that it would take speakers longer to produce appropriately specified descriptions in the *Low Distortion* condition. We also included two predictions of secondary importance: (1) that speakers would be more likely to overspecify than to underspecify targets; and (2) that speakers would gaze at fewer non-target items prior to speech onset in the *Low Distortion* condition. In the end, we opted for a more comprehensive time-series approach to analyzing the eye data, and so the eye-tracking results for the pre-registered hypothesis are reported as part of the online materials.

Method

A link to the pre-registration for this experiment can be found in the OSF repository.

Participants

We collected data from a a total of 47 participants, with data from 11 excluded for reasons specified below, leaving 36 participants included in the analysis. The participants were 24 women and 12 men. The sample size of 36 was determined in advance, through a power analysis based on a small pilot study (80% power for the effect observed in the pilot). All participants were recruited from the University of Glasgow campus and were either paid £6 or received course credit.

From our pilot study, we were concerned that some speakers would opt for a strategy of overdescription—that is, always using a size modifier even when there was not a competitor letter in the display. The problem with this behavior is that on test trials in the *Singleton-Contrast* condition, speakers could simply continue using the modified description, which would then spuriously appear to be appropriately specified. We pre-registered our intent to exclude any participants who inappropriately used size modifiers on more than half of the final training trials for each sequence in the *Singleton-Contrast* level. This resulted in the exclusion of ten participants. One additional participant was replaced due to a problem that we did not anticipate in our pre-registration protocol. This speaker used excessively long descriptions on each trial, seeming to have misconstrued the task as one of providing fine-grained description rather than providing sufficient information for the experimenter to locate the target.

Subjects gave written informed consent before beginning the experiment and were fully debriefed after the experiment had finished. Our procedures fully complied with the ethical code of conduct of the British Psychological Association.

Experimental Setup and Task

The experiment was interactive with the participant playing the role of speaker and the experimenter playing the role of the listener. The speaker and the listener sat in different areas of the testing room and looked at separate computer monitors throughout the experiment. Both were seated facing in opposite directions so that they were unable to see the other's display. In each trial, the speaker was asked to describe a highlighted target letter, which appeared on their monitor, to the listener. The listener then identified this letter on his own screen and selected it using a computer mouse. The target letter appeared on the speaker's screen within a grid among other distractor letters (Figure 1). The speaker was informed that in each trial the listener would have the same letters on their monitor but that they may be arranged in a different format compared to the grid that appeared on their screen.

Design

There were two factors in the design, *Distortion* (*Low*, *High*) and *Shift Direction* (*Singleton-Contrast* and *Contrast-Singleton*), forming a full-factorial 2x2 within-participant design.

Materials

Each display consisted of a five-by-four grid containing uppercase letters (A-Z) of different font size and color. All letters appeared in uppercase Arial font. The font sizes for targets and competitors/foils were randomly generated for each trial, with 'small' defined as between 64–96pts and 'large' as 32pts higher than the smaller letter in a pair. We refer to a single 'sequence' as the set of training trials and the single test trial associated with a single target-competitor-foil triad. There were 48 sequences in each experimental session. Displays and target/foil pairs were randomly generated for each participant. For each sequence, the number of training trials was randomly chosen from a range of 6 to 9. Given these parameters, each experimental session could have contained anywhere between 336 trials (6×48 training trials plus 48 test trials) and 480 trials (9×48 training trials plus 48 test trials).

Each sequence for each session was based on a randomly generated original "prototype" display, which was used as the test trial, with the training displays generated as distortions of this prototype. The identity, color, and size of the target letter in each sequence were fixed across all displays. The identity of the target letter was chosen randomly for each sequence in each session, with the constraint that the same letter could not be used as target more than once per session within each block of 24 sequences formed by the *Distortion* factor. After the selection of the target for a given sequence, the foil letter was selected from the remaining set of letters, with the probability of selection inversely proportional to its similarity to the target, as derived by norms given in Simpson et al. (2013). By biasing the selection toward visually similar letters, we attempted to increase the likelihood that speakers would fail to detect the difference between a letter with the same identity (e.g., target 'O', competitor 'Q'). The random selection process also meant that each participant would get mostly distinct letter pairs, which allows us to treat items as a fixed effect in our analyses (Clark, 1973).

In addition to the target and competitor/foil, there were three sets of distractor letters scattered among some of the remaining squares in the grid. The distractor letters were randomly chosen from the set of letters excluding the target and competitor. Each set in each sequence had letters of a different color, each randomly chosen (without replacement) from a palette of ten colors. The first set was of the same color as the target and competitor, and had either four or five letters. The second set was of a different color and also had either four or five letters. The third set was also of a different color and had one or two letters. The sizes of the distractor letters were randomly chosen from within the range of 64–128 pts.

Next, the letters for each prototype were assigned positions within the display. For a given sequence, the target and competitor (or foil) letters always appeared in the same colors and with fixed positions across all training and test displays. The assignment of the target and competitor (or foil) positions was random, with the constraint that the two letters must be at least four spaces apart using a city-block metric. The positions of the distractor letters were assigned randomly.

The training trials for each sequence were created by distorting the prototype. *Low Distortion* displays were created by randomly selecting either two or three distractor letters from the prototype and moving them to an adjacent empty space in the grid. Any letter that was "locked in" (i.e., all surrounding spaces occupied) was never selected to move. For *High Distortion* patterns, the positions of all of the distractor letters were randomly reassigned to any space not occupied by the target or competitor/foil, and the colors of any two of the distractor sets could be swapped.

In each speaker display, the target was highlighted with a yellow surrounding square. There was no such indication on the listener's displays. We wanted to make it more difficult for speakers to identify the competitor letter using peripheral vision. To this end, we added a slight Gaussian blur to the speakers' images using the `convert` command within the ImageMagick suite of command-line tools (version 8:6.7.7.1, http://www.imagemagick.org), with the sigma parameter set to 8 and radius set to 0 (0x8).

The listener's displays were created by simply randomizing the positions of the letters in the speaker's grids. Thus, while the locations of the target and competitor/foil of each sequence were fixed for the speaker, they varied from trial to trial for the listener. For information about the sequencing of trials within a block, please see the supplementary information provided in our OSF repository.

Of the 48 test trials presented in each experimental session, 24 were in the *Low Distortion* condition (with 12 in the *Singleton-Contrast* condition, and 12 in the *Contrast-Singleton* condition) and 24 were in the *High Distortion* condition (12 in the *Singleton-Contrast* condition, 12 in the *Contrast-Singleton* condition).

Apparatus

The experimental stimuli were presented on a 19" LCD Dell desktop computer monitor (4:3 aspect ratio, resolution 1024 pixels wide by 768 pixels high). Participants were seated 45–55 cm away from the monitor. A microphone was placed above the participant's computer monitor to record their referential descriptions. Eye movements were recorded using an Eyelink 1000 (SR Research) remote eye tracker, with a sampling rate of 500Hz.

Procedure

At the start of any given trial, an empty grid appeared on the speaker's screen, with the yellow square marking the location where the target would appear. After one second, the preview screen was replaced with the full display. Audio recording of the speaker's response began simultaneously with the presentation of the full display. The trial ended when the listener selected the object designated by the speaker. The speaker could not see the listener's screen or mouse pointer, and received no feedback regarding whether or not the listener had selected the intended referent. If the speaker failed to provide sufficient information to identify the target, the listener asked the speaker for clarification (e.g., "Which 'W' do you mean?"). Any such clarification exchanges appeared in the audio recording for the trial and were noted during later transcription.

Data Analysis

Our analysis focused on three categories of measurements: (1) speech content; in particular, use of a size modifier (e.g., large/small); (2) speech onset latency, defined as the time taken to produce the first content word as measured from the onset of the display; and (3) eye movements.

For each of the 48 sequences for each speaker, we transcribed and coded the audio recordings for two trials: (1) the last trial of the training sequence; and (2) the test trial. The last training trial was needed in order to provide baseline data for the speech onset latency in the test trial, and to identify test trials to be excluded (see Results and pre-registration document).

For misspecification rate, we coded whether or not a size modifier was used by the speaker in the test trial (Table 1). Misspecifications were determined from these codes as follows. In the *Singleton-Contrast* condition, which required a modifier, the codes *NO*, *AS*, and *AO* were counted as misspecifications. In the *Contrast-Singleton* condition, where a modifier was not required, all codes other than *NO* were counted as misspecifications.

Onset times of utterances were identified and entered into a data table in milliseconds. The following criteria were applied when identifying utterance

Table 1: Coding of speech utterance types.

Category	Description	Example
NO	No size modifier	"W", "the W", "the red W"
PR	Pre-nominal size modifier	"small W", "large W"
PO	Post-nominal size modifier	"W that is small", "W, big"
DE	Deleted adjective	"sm— uh just the W"
AS	Addition by self-repair	"W… big W"
AO	Addition due to other-repair	"W" ("which one"?) "big W"

onsets:

1. Trials were discarded if the speech was unidentifiable.
2. Any filled pauses or articles were ignored (um, uh, the); speech onset was identified as the first content word (e.g., adjective or noun), even if the adjective referred to color rather than size (e.g., for "uh… the blue W" onset was taken to be at the onset of the word "blue").
3. If speakers corrected themselves after an error (e.g. "pink W…eh sorry blue W") onset of the correction (i.e. "blue") was recorded. However, such repaired utterances were not used in the analysis of speech onset.

For all appropriately-specified descriptions, we counted up all non-target fixations (with a minimum fixation duration of 100ms) that took place prior to speech onset and tested the effect of *Distortion* on fixation counts. We predicted a higher rate of pre-onset fixations in the *Low Distortion* condition, based on the rationale that speakers would experience a weaker memory signal for the entrained description and would thus engage in more checking of context during speech planning.

Results and Discussion

We performed all statistical analyses using the R statistical programming environment, version 3.3.3 (R Core Team, 2017). Linear mixed-effects models were estimated using lme4 package version 1.1.21 (Bates et al., 2015b). We sought to include the maximal random effects structure justified by the design (Barr et al., 2013), which entails by-subject random intercepts and by-subject random slopes for both main effects (*Distortion* and *Shift Direction*) and their interaction. It was not necessary to include item as a random factor since the displays were randomized and specific target/foil pairs defined separately for each participant (Clark, 1973). We derived p-values using the t-to-z heuristic, which enabled us to perform pre-specified one-tailed tests where required. Unless otherwise noted, tests were two-tailed with $\alpha = .05$. *Shift Direction* was coded as *Contrast-Singleton* = -.5, *Singleton-Contrast* = .5, while *Distortion* was coded as *High* = -.5, *Low* = .5.

Misspecification Rate

The 36 participants included in the analysis completed a total of 1728 trials, 1677 of which were used in the analysis. The remaining 51 were excluded, 41 because in the last training trial prior to the test trial, participants did not use a modifier even though it was required, and ten because poor recording quality made it difficult to transcribe the speech. The data are shown in Figure 3.

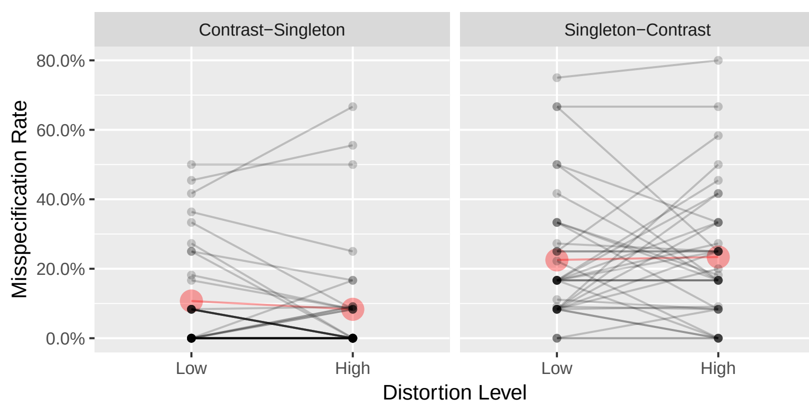

Figure 3: Misspecification rate by Distortion Level and Shift Direction. Connected black points are individual participants and red points are grand means.

Table 2 shows strong differences in the distribution of utterance types across the levels of *Shift Direction*, but little evidence for any effect of *Distortion*. When speakers had entrained on modified nouns and were tested in a context requiring a bare noun, speakers only overspecified about 9.3% of the time overall. Speakers were far more likely to misspecify references when they had entrained on bare nouns and the test context required a modifier. About 23.1% of the

Table 2: Distribution of utterance types across condition, Experiment 1. NO = no misspecification; PR = pre-nominal modifier; PO = post-nominal modifier; DE = deleted adjective; AS = Addition by self-repair; AO = addition by other-repair. For Contrast-Singleton, PR, PO, DE, AS, and AR were misspecifications; for Singleton-Contrast, it was NO, DE, AS, and AO.

Shift Direction	Distortion	NO	PR	PO	DE	AS	AO
Contrast-Singleton	Low	89.4%	6.4%	1.7%	1.9%	0.7%	0.0%
Contrast-Singleton	High	92.0%	3.9%	1.5%	2.7%	0.0%	0.0%
Singleton-Contrast	Low	1.9%	58.4%	18.8%	3.3%	14.3%	3.3%
Singleton-Contrast	High	1.7%	58.6%	17.5%	4.1%	14.4%	3.6%

time speakers failed to include a modifier in the first instance (pre-nominally or post-nominally). Typically, if a modifier was included, it was included as a self-repair ("the W… uh small W").

For the inferential analysis of misspecification rate, we performed logistic regression using `glmer()`. The logistic regression model converged with maximal random effects, but reported singularity in the variance-covariance matrix. Given doubt as to the interpretation of singular models (Bates et al., 2015a), we fit a second model in which we reduced the random effects structure until the singularity was removed. This second model included a random intercept and random slopes for *Shift Direction* and the interaction term, but not for *Distortion*. We report results from the second model.

There was little evidence to support the main prediction of a main effect of *Distortion*. Misspecifications were observed on 16.6% of trials in the *Low Distortion* condition, compared to 15.8% in the *High Distortion* condition. This difference was not significant (pre-registered one-tailed test), $\beta = 0.14$, $SE = 0.16$, Wald $z = 0.87$, $p = 0.191$. There was also little evidence for an interaction between *Shift Direction* and *Distortion*, $\beta = -0.41$, $SE = 0.37$, Wald $z = -1.11$, $p = 0.267$. The effect of *Shift Direction*, in contrast, was significant, $\beta = 1.47$, $SE = 0.32$, Wald $z = 4.53$, $p < .001$, with a higher rate of misspecification in the Singleton-Contrast condition (as noted above, 23.1% versus 9.3%). This implies that speakers were more likely to underspecify targets (e.g., calling the smaller of two Ws 'the W') than to overspecify them (e.g., calling the lone W 'the small W').

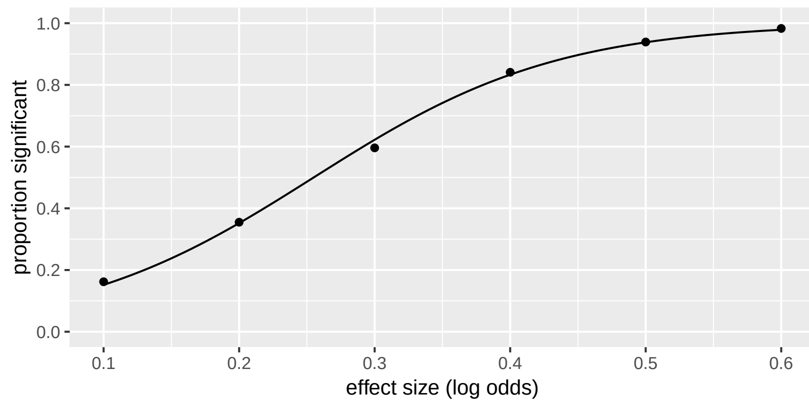

Figure 4: Sensitivity analysis for Experiment 1. The points are the proportion of significant tests at each of the six effect sizes examined; the line is a fit from a logistic regression model to allow smooth interpolation.

The above results suggest little evidence for a systematic effect of *Distortion*; however, it is possible that participants were sensitive to the manipulation, despite not showing an overall consistent effect in the predicted duration. In other words, perhaps there are individual differences across participants in how memory affects referential description, but no consistent overall effect. To test this hypothesis, we ran a null-hypothesis significance test on the random slope parameter for *Distortion*. The estimated random slope of 0.20 logits was not significantly different from zero, $\chi^2(1) = 0.06$, $p = 0.814$.

How sensitive was our experiment to rejecting a null effect of the critical *Distortion* variable? To address this question, we ran a sensitivity analysis based on the parameter estimates obtained from the linear mixed-effects analysis. A sensitivity analysis gives insight into the power of a null-hypothesis test over a range of effect sizes, characterizing the severity of the test (Mayo, 2018). For this analysis (which was not pre-registered), we varied the fixed effect of distortion over six steps on the logit scale (.1 to .6, in steps of .1). For each of these parameter values, we simulated 1,000 datasets, deriving all other necessary parameters from the model estimates. Each dataset was analyzed using a linear-mixed effects model with a by-subject random intercept and a by-subject random slope for the main effect of *Distortion*. We estimated power at each effect size by calculating the proportion of significant tests (Figure 4).

The sensitivity analysis suggests that the test of the main effect of *Distortion* had: (1) 80% power to detect a significant effect of at least 0.38 logits, corresponding to a effect on misspecification rate of about 2.9% between conditions (e.g., 15.7% in the high distortion condition versus 18.6% in the low distortion condition); (2) 90% power to detect an effect of at least 0.45 logits, or a difference of about 4.2% (e.g., 14.9% versus 19.1%); and (3) 95% power to detect an effect of at least 0.52 logits, or a difference of about 5.3% (e.g., 14.2% versus 19.5%).

Speech onset latency

Our second main prediction concerned the differential speech onset latency for appropriately specified descriptions. Our prediction was that speakers would be less likely to shift from the entrained description to a more contextually appropriate description in the *Low Distortion* condition than in the *High Distortion* condition, due to a stronger retrieval of the entrained response. For this analysis, in addition to the 51 trials excluded for reasons detailed above, a further 272 trials were excluded where speakers misspecified the target, and two more where the speech onset could not be determined.

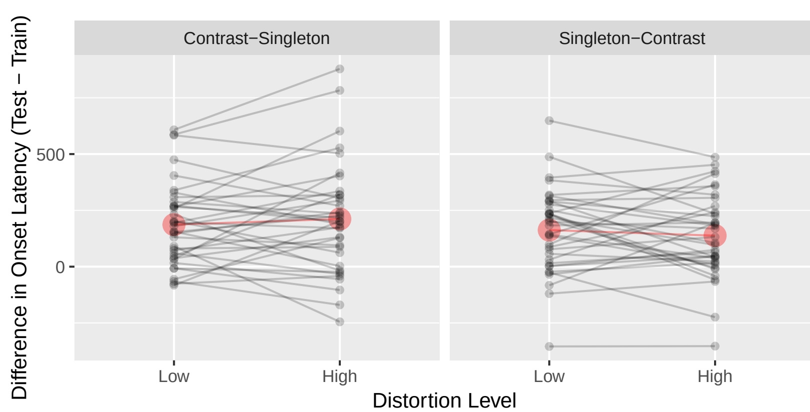

Figure 5: Change in speech onset latency from training to test by Shift Direction and Distortion Level. Positive values indicate higher latencies at test; connected black points are individual participants and red points are grand means.

The data are shown in Figure 5. Parameters were estimated using the `lmer()` function under maximum likelihood with identity link and Gaussian variance. The dependent variable was the speech latency for the test trial minus the speech latency for the final training trial for that sequence; in other words, the change in speech latency incurred by abandoning the entrained description. We tested against the null using a two-tailed test on the Wald z statistic with $\alpha = .05$.

We once again encountered singularity when fitting the maximal random-effects model, so we fit a second model with reduced random effects (the converging non-singular model had random intercepts and random slopes for *Shift Direction*). Again, we report the results from the non-singular model.

The main prediction that speakers would encounter greater difficulty producing appropriately specified descriptions in the *Low Distortion* condition was not supported: there was only a difference of 5 ms between the Low and High conditions, with means of $M = 175$ ms ($SD = 421$) and $M = 180$ ms ($SD = 444$) respectively, and a non-significant main effect of *Distortion*, $\beta = 0.63$, $SE = 21.48$, Wald $z = 0.03$, $p = 0.977$. There was also no significant main effect of *Shift Direction*, $\beta = -53.17$, $SE = 47.94$, Wald $z = -1.11$, $p = 0.267$. Finally, the interaction was also not significant, $\beta = 57.24$, $SE = 43.13$, Wald $z = 1.33$, $p = 0.184$.

Eye gaze

The modest rate of misspecification in the experiment indicates reliance on memory, but we found no clear evidence for the main prediction of stronger retrieval of entrained descriptions in the *Low Distortion* condition. This result is ambiguous: it could be taken as support for the idea that speech content is not strongly influenced by ordinary memory processes, but only if the manipulation of layout was successful in inducing stronger memory associations in the *Low Distortion* condition.

To verify this, we plotted the probability of gazing at various types of images over time (Figure 6). There is some suggestion that the memory manipulation was effective in the predicted direction, but the effect appears small. The figure

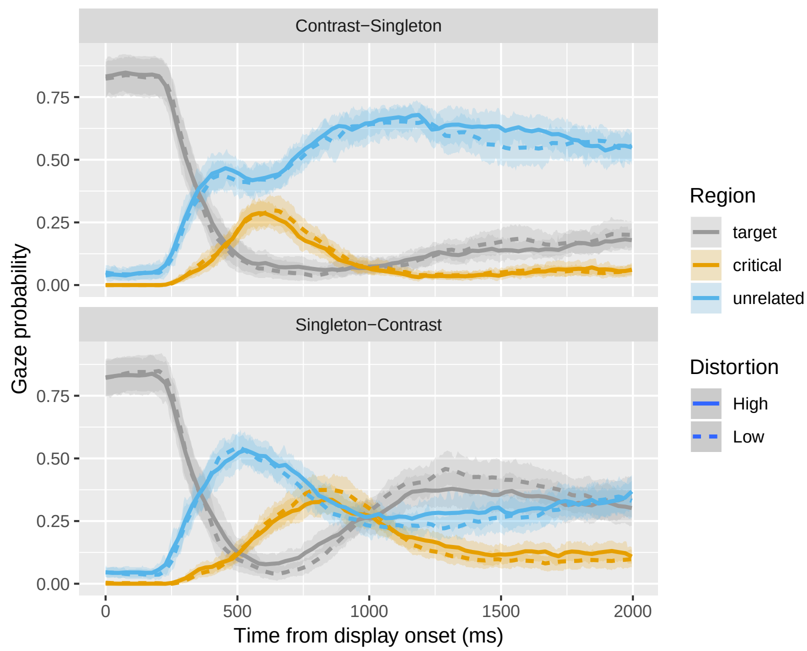

Figure 6: Gaze probability by Shift Direction and Distortion. The critical image was a competitor in the Singleton-Contrast condition and a foil in the Contrast-Singleton condition. The shaded region around each line represents the 95% confidence interval obtained by bootstrapping subjects.

indicates that speakers typically started the trial already fixated on the pre-cued target image and began looking away about 250 ms after display onset. Speakers appear to look away from the target and toward the critical image (competitor or foil) slightly more rapidly in the *Low Distortion* condition. Although the pattern is consistent with stronger memory effects in the *Low Distortion* condition, the size of the effect is not such to fully rule out the possibility that the memory manipulation was too weak.

We also carried out our pre-registered test of a difference in the rate of non-target fixations prior to speech onset for appropriately-specified descriptions. We predicted a higher fixation rate for the *Low Distortion* condition. We fit a generalized linear mixed effects model to the count data with a log link and Poisson variance function. The model included a by-subjects random intercept and a random slope for the effect of *Distortion*. The covariance matrix for random effects was singular, so we dropped the random slope and re-fit the model.

No statistically significant difference was detected (two-tailed) between the mean fixation rates of 1.45 (SD = 0.63) for the *Low Distortion* condition, and 1.44 (SD = 0.62) for the *High Distortion* condition, $\beta = 0.01$, $SE = 0.04$, Wald $z = 0.15$, $p = 0.882$.

Experiment 2

Speakers reliably misspecified referents in Experiment 1, which unequivocally demonstrates the involvement of memory processes—it would be unlikely for a speaker to call an average-sized lone W 'the small W'—in the absence of the encoding experience from the training trials where it appeared with a larger one. However, the non-verbal eye gaze measurements suggest that the main manipulation may have been too weak. Experiment 2 included several improvements to increase training-test similarity and strengthen the memory associations with the target descriptions.

It is possible that memory effects in Experiment 1 were obscured by the high similarity among memory representations across stimulus items. All of

Figure 7: Example training and test displays, Experiment 2. The target was indicated to the speaker by a green square surrounding the image.

the target items were letters of the alphabet, and the only type of modification ever required was a size modification (e.g., "the small W"). This would have created high overlap among memory representations that could have led to interference during retrieval. Another potential problem with Experiment 1 is that entraining on gradable adjectives such as "small" or "large" may have been ineffective for memory encoding since the meanings of these words are so context dependent.

We addressed these limitations using a variety of everyday objects as targets and competitors in Experiment 2 (Figure 7), similar to the type of stimuli used in Gann and Barr (2014). For each pair, the target was always a very typical member of the category (moreso than the competitor), such that when described in a context by itself there would be a strong tendency to use a basic level term (e.g., the candle, the apple, the car) (Rosch et al., 1976). Competitors were chosen so as to elicit target modifiers that would be very unlikely to occur in the absence of the target. For the candle example, the competitor was a highly similar half-melted candle, such that a typical description of the target would be "the unmelted candle." We chose foil objects that were visually similar to the competitor but from a different category of object. For a full list of target-competitor-foil triplets, see the supplementary materials in the data repository.

The above manipulation also addresses an additional weakness of Experiment 1, which was the high rate of data loss due to speakers who opted for a strategy of always including a size modifier, whether or not it was needed. In this experiment, speakers would need to focus on different dimensions for each category of object, which was intended to discourage this strategy and reduce data loss.

A second way we sought to strengthen memory effects was by removing pre-cuing of the target location. In the previous experiment, before the target or any other images appeared, speakers were directed toward the target's location. Since the location itself might operate as a memory cue, retrieval processes may have already begun prior to display onset, making their effects less detectable when they were measured at a later point. To remedy this, in the current experiment, the highlight indicating the target location occurred simultaneously with display onset.

Finally, instead of creating low and high distortions of a single spatial arrangement prototype, we simply manipulated whether the arrangement at test was identical to the training arrangement (*Congruent* condition), or a different random arrangement (*Incongruent* condition), forming the factor of *Congruency* (Figure 7). Note that for the *Incongruent* displays, the target location remained fixed over training and test displays, while the locations of all other images (including the competitor/foil) were randomized.

Method

The method of Experiment 2 was similar to Experiment 1, and so we only describe the differences. A link to the pre-registration for this experiment can be found in the OSF repository.

Participants

We collected data from a a total of 37 University of Glasgow students (24 women and 13 men), with data from one participant excluded due to overdescribing in more than 50% of the last training trials before the corresponding test trial. All participants were paid £6 or received course credit.

Design

There were two factors in the design, *Congruency* (*Congruent* versus *Incongruent*) and *Shift Direction* (*Singleton-Contrast* and *Contrast-Singleton*), forming a full-factorial 2x2 within-participant design. Both factors were also manipulated within each stimulus set.

Materials

Each display consisted of a five-by-four grid containing various types of everyday objects (see Figure 7). The experiment contained 48 "sequences" of trials, each consisting of a number of training trials followed by a single test trial (with each sequence being defined as the collection of training and test trials all associated with a single target/competitor/foil triplet). Each triplet appeared an equal number of times in all four conditions of the 2x2 design, counterbalanced across participants using stimulus lists.

For each sequence, the number of training trials was randomly selected, with a range from 6 to 9. Given these parameters, each experimental session could have contained between 336 (7 x 48) and 480 (10 x 48) trials. For each stimulus set, 7 to 10 additional images unrelated to the target were randomly chosen from a database of stimulus images. Images were re-used across trials within a sequence, but not across different stimulus sets. The displays were checked manually by two lab assistants to ensure that the unrelated items were sufficiently dissimilar to the target so as not to influence descriptions of the target.

Target and competitor items were normed beforehand by 68 Native English speaking volunteers using the web-based survey platform SurveyMonkey. A number of items were updated or replaced based on our norming feedback. Four entirely new stimuli pairs were added to our original list (for a complete list of the Target and Competitor objects used please see the supplementary information provided in the OSF repository).

Apparatus

The apparatus was identical to Experiment 1, with the exception that the eyetracking sampling rate for all participants was set to 250 Hz.

Procedure

The procedure was identical to the previous experiment, with the exception that on each trial, the cue for the target location appeared simultaneously with the rest of the display. Additionally, although speakers and listeners had different arrangements of each set of images within the grid, the listener's arrangement was held constant within each sequence to facilitate easier identification of the target. Our rationale was that with predictable target locations, listeners would be faster to identify targets during training, which might lead the speaker to entrain more strongly on the referential precedent.

Data Analysis

The measurements and predictions were identical to Experiment 1, with the difference that all mixed-effects models also included by-item random intercepts and slopes, since items (sets of target/competitor/foil/unrelated images) repeated across participants and were likely to induce different patterns of modification.

Occasionally a speaker would use a single subordinate-level term to distinguish the target from the competitor (e.g., "notes" instead of "paper money" to distinguish a stack of notes from a pile of coins). In these instances, we coded the speech as category *PR*. Cases where speakers provided disambiguating information after as well as before the head noun ("the woolen gloves that are red"), were coded as *PR*, as long as the information before the head noun seemed sufficient to disambiguate the target from the competitor. Thus, the choice of *PR* versus *PO* captures whether any adaptations took place up to (and including) the head noun, or somewhere after (*PO*).

Unlike Experiment 1, we only coded speech onset times for appropriately-specified descriptions. Another difference was that we also established exclusion criteria for items (stimulus sets). We considered the last training trial for each participant on each item on which the critical object was a foil, and removed items where speakers used a modifier that would have distinguished the target from the (absent) competitor. Any items where the rate of modifier use was greater than 50% was removed.

Results and Discussion

Of the 1728 test trial observations we recorded, we removed 252 from seven stimulus items that met our exclusion criterion (see above), and an additional 176 observations where either speakers had used a modifier in the last training trial where it was not appropriate (143), or the response could not be identified from the sound recording (33). This left 1300 trials remaining for the analyses below.

Misspecification rate

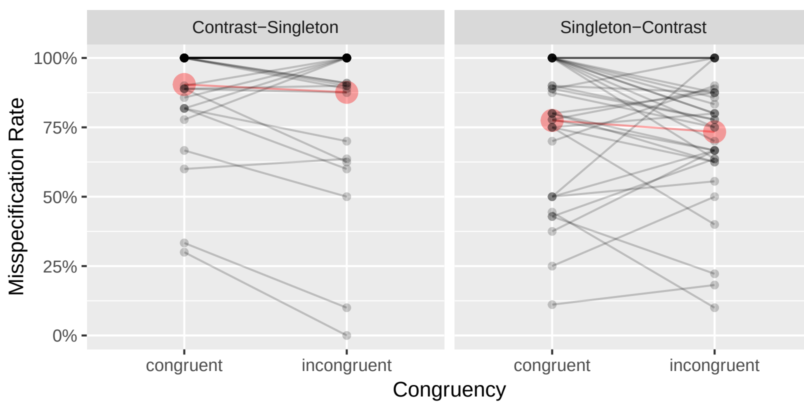

Figure 8: Misspecification rate by Congruency and Shift Direction. Connected black points are individual participants and red points are grand means.

Table 3: Distribution of utterance types across condition, Experiment 2. NO = no misspecification; PR = pre-nominal modifier; PO = post-nominal modifier; DE = deleted adjective; AS = Addition by self-repair; AO = addition by other-repair. For Contrast-Singleton, PR, PO, DE, AS, and AR were misspecifications; for Singleton-Contrast, it was NO, DE, AS, and AO.

Shift Direction	Congruency	NO	PR	PO	DE	AS	AO
Contrast-Singleton	Congruent	9.5%	76.4%	12.1%	1.7%	0.3%	0.0%
Contrast-Singleton	Incongruent	12.4%	75.4%	10.7%	1.2%	0.3%	0.0%
Singleton-Contrast	Congruent	3.4%	12.4%	8.4%	0.3%	22.5%	53.0%
Singleton-Contrast	Incongruent	2.7%	21.4%	6.1%	1.0%	21.1%	47.6%

Table 3 shows strong differences in the distribution of utterance types across the levels of *Shift Direction*, but little evidence for any effect of *Congruency*.

When speakers entrained on modified nouns, overspecification at test occurred in about 89.1% of cases. When they entrained on bare nouns, 74.5% of cases failed to include a modifier in the first instance; typically, if a modifier was included, it was included as a self-repair.

For the misspecification rate analysis, we performed logistic regression using `glmer()`. The logistic regression model of misspecification rate converged with maximal random effects, but reported singularity in the variance-covariance matrices. We fit a second model in which we reduced the random effects structure until the singularity was removed. The model that converged included by-subject and by-item random intercepts; by-subject random slopes for *Shift Direction*, *Congruency*, and their interaction; and by-item random slopes for *Shift Direction* and *Congruency* but not for the interaction. All covariance parameters were constrained to zero. We report the results from the second model.

There was some evidence for the main prediction: misspecifications were observed on 84.5% of trials in the *Congruent* condition, compared to 80.1% in the *Incongruent* condition, a significant main effect of *Congruency*, (pre-registered one-tailed test), $\beta = 0.42$, $SE = 0.22$, Wald $z = 1.88$, $p = 0.030$. There was little evidence for an interaction between *Shift Direction* and *Congruency*, $\beta = -0.10$, $SE = 0.41$, Wald $z = -0.25$, $p = 0.806$. The effect of *Shift Direction*, in contrast, was significant, $\beta = -1.71$, $SE = 0.38$, Wald $z = -4.56$, $p < .001$, with higher rates of misspecification in the Contrast-Singleton case (as mentioned above, 89.1% versus 74.5%). In other words, in this experiment, speakers were more likely to overspecify than underspecify.

Following up on this significant effect of *Congruency*, we also examined whether there were significant individual differences across participants in the magnitude of the effect by testing the significance of the corresponding random slope (this analysis was not pre-registered). However, we failed to detect any differences across subjects or items. The estimated by-subject random slope of 0.26 was not significantly different from zero, $\chi^2(1) = 0.09$, $p = 0.770$. The test of the by-item random slope (estimate = 0.42) was also not significant,

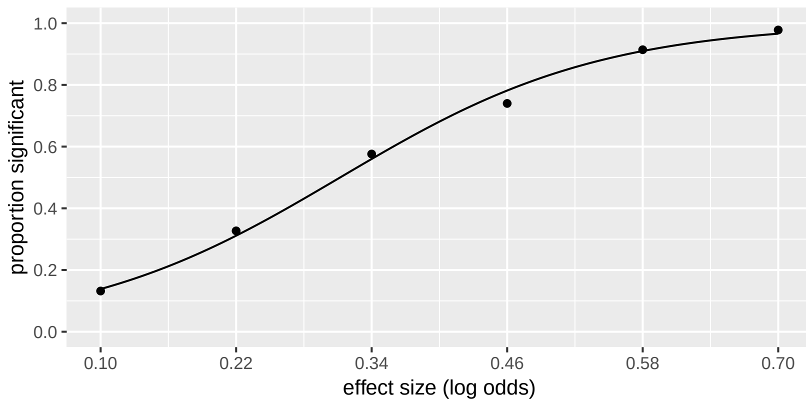

Figure 9: Sensitivity analysis for Experiment 2. The points are the proportion of significant tests at each of the six effect sizes examined; the line is a fit from a logistic regression model to allow smooth interpolation.

$\chi^2(1) = 0.39$, $p = 0.533$.

We also performed a sensitivity analysis for the test of the main effect of *Congruency*, following the same procedure described for Experiment 1, with the difference that the effect size range was from .1 to .7 logits. (This analysis was also not pre-registered.) Results are shown in Figure 9. The sensitivity analysis suggests our test of the main effect of *Congruency* had: (1) 80% power to detect an effect of at least 0.47 logits, corresponding to an effect on misspecification rate of about 4.6% between conditions (e.g., 77.0% in the incongruent condition versus 81.6% in the congruent condition); (2) 90% power to detect an effect of at least 0.57 logits, or a difference of about 5.1% (e.g., 77.2% versus 82.3%); and (3) 95% power to detect an effect of at least 0.65 logits, or a difference of about 5.4% (e.g., 77.2% versus 82.6%).

Speech onset latency

As in the previous experiment, the prediction was that speakers would have more difficulty shifting to an appropriately specified test description when memory associations were stronger (i.e., in the *Congruent* condition). The means in the *Congruent* and *Incongruent* conditions were inconsistent with this prediction, $M = 2277$ ms ($SD = 975$) and $M = 2481$ ms ($SD = 981$) respectively. However, the very high rate of misspecification in the current experiment (82.3%) left very few appropriately-specified observations for analysis (only 230). Given the very small number of remaining observations, we opted to forgo any further statistical analysis. Because we only had onset data for a small minority of observations, we also did not pursue our pre-registered analysis of the rate of pre-onset fixations across congruency conditions.

Eye gaze

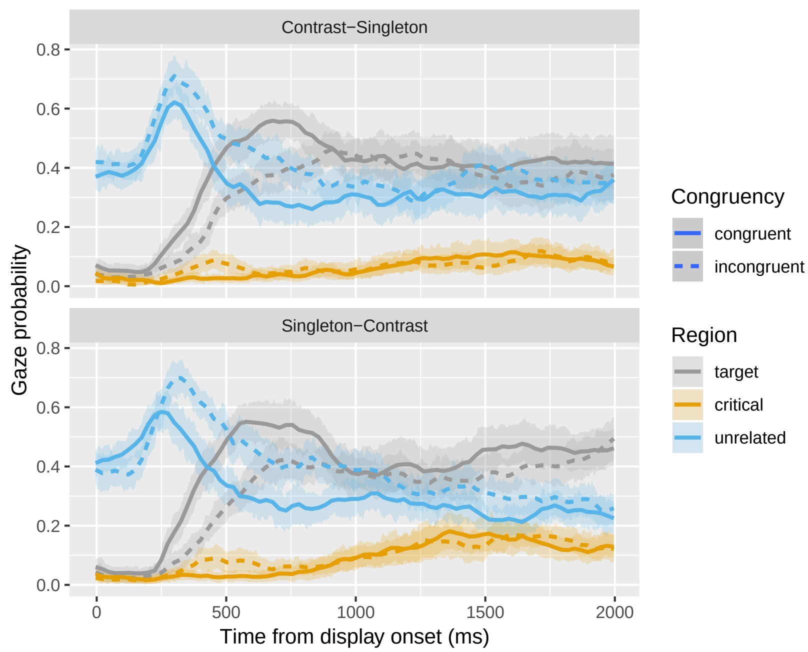

Figure 10: Gaze probability by Shift Direction and Congruency. The critical image was a competitor in the Singleton-Contrast condition and a foil in the Contrast-Singleton condition.

In contrast to Experiment 1, a plot of the probability of eye gaze on the various regions (Figure 10) indicates unequivocal memory effects: when the display arrangement was congruent at test, speakers looked away from the target and toward the critical object much more rapidly.

Unlike the previous experiment, the main prediction did receive some statistical support: speakers relied on remembered descriptions significantly more often when the arrangement of the test display was more similar to the arrangement at training. However, the effect of similarity was small, corresponding to about a 4.5% difference. In contrast, the differences in eye movements were quite strong: for instance, the rate of fixating on the target 500 ms after onset was about 50% in the *Incongruent* condition versus 30% in the *Congruent* condition. So although perceptual aspects of the conversational situation are clearly stored and affect processing, these aspects may only have very weak effects on speech content.

One notable difference from Experiment 1 was the very high overall misspecification rate, which was about 82.3%, compared to about 16.2% in Experiment 1. What might account for this difference? This might be attributable to semantic/pragmatic differences between size modifiers and other types of modifiers. Perhaps because size modifiers have more relational semantics than other types of modifiers (Grodner and Sedivy, 2011), they are used in a more context-specific manner. Another possibility is that the constant use of size modification in Experiment 1 led speakers to pay more attention overall to the presence or absence of a size contrast in any given display.

Experiment 3

In Experiments 1 and 2, we used spatial arrangement as a memory cue. While spatial cues allow convenient measurement of implicit aspects of memory processes through eyetracking, they have the disadvantage of potentially low ecological relevance for communication. Although the eye movement data showed that speakers did indeed store information about spatial configuration, spatial information is not something that speakers need to regularly attend to for their references to succeed. The perceptual information associated with conversational partners provides a far more important and ecologically relevant set of cues.

When language users interact, particularly in a face-to-face setting, it would seem likely that they would develop links between the content of the dialogue and perceptual features of their interactions, such as how their interlocutors look and sound. The ordinary memory view assumes that these perceptual features can drive a resonance process that makes relevant information from past conversations readily accessible (Horton and Gerrig, 2005). We designed Experiment 3 around this key assumption.

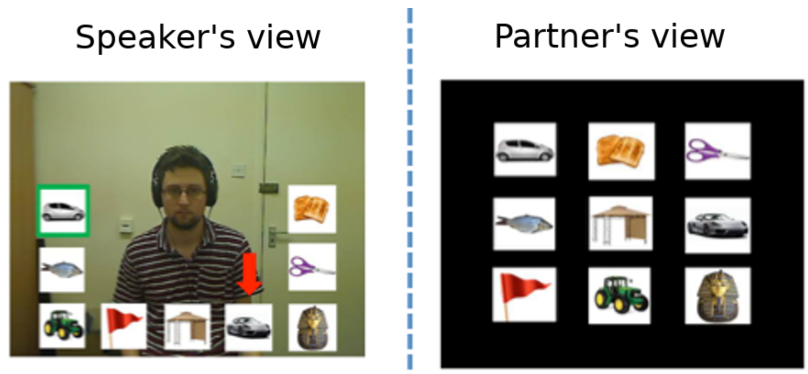

Figure 11: Example displays for the speaker and the confederate partner, Experiment 3. (Note that the red arrow, which indicates the location of the competitor, is included for expository convenience, but did not appear on the speaker's display.)

Sighted language users in Western cultures generally look at their addressees while communicating. As a result, the perceptual experience of seeing one's partner is confounded with the speaker's mental representation of that person's conversational role as an addressee. For the purpose of our study, it was necessary to deconfound these two streams of information to be able to measure memory effects independently from effects about the speaker's beliefs their common ground with the addressee. We did this by having speakers communicate with two different partners over independent video and audio links. Both of the partners were together, but in a separate room from the speaker. The two partners (who were experimenters) alternated in their role as addressee. The video link showed one of the two partners at a time. We manipulated which partner appeared on the speaker's screen independently from which partner had access to the audio link relaying the speaker's live speech. This setup makes it possible for speakers to be looking at one partner who cannot hear them while addressing another, unseen partner.

During training phases, speakers developed memory associations between targets and expressions while addressing and viewing one of two partners. Both partners wore headphones; during training, the on-screen partner could hear and respond to descriptions via audio link, while the off-screen partner wore a blindfold and heard masking noise to limit their access to the exchange. The images that the speaker conversed about appeared superimposed over the image of the partner (see Figure 11). The objects depicted in the images were similar to the ones we used in Experiment 2 and, as in that experiment, speakers also entrained on either modified or unmodified descriptions. Note that during training, the visible partner and the addressee were one and the same.

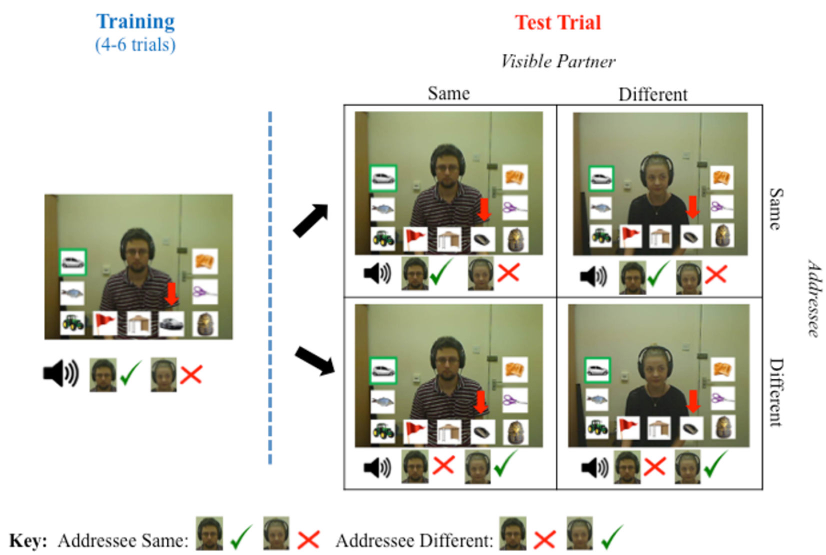

Figure 12: Design of Experiment 3, showing one training and one test trial for an example item in the Contrast-Singleton condition. The four versions of the test trial correspond to all four conditions obtained by factorially combining Visible Partner and Addressee. The red arrow indicates the competitor/foil object for expository purposes, and was not shown on the speaker's display. The loudspeaker icon indicates which partner was the addressee (i.e., could hear the speech) on that trial in that condition.

As in the previous experiment, test trials differed from training trials in the identity of a critical object: in the *Singleton-Contrast* condition, the foil from training became a competitor at test (potentially inducing underspecification of the target), while in the *Contrast-Singleton* condition, the competitor at training became a foil at test (potentially inducing overspecification of the target). We also independently manipulated the congruence of the test situation with the training situation along two separate dimensions: *Addressee* (same or different) and *Visible Partner* (same or different; see Figure 12). The critical question was whether speakers would be more likely to misspecify referents at test when the visible partner was the same as the one they had seen (and spoken to) during training, compared to when it was the partner they had not spoken to. If partners serve as memory cues, speakers should be more likely to misspecify referents at test when they are looking at the partner they spoke to during training (regardless of whether they are currently addressing this partner). To maximize power, we pre-registered a one-tailed test of this prediction (greater misspecification in the *Visible Partner: Same* condition). Additionally, the inclusion of the *Addressee* factor makes it possible to test the extent to which misspecified descriptions are the result of speakers using common ground—that is, of them preferring to continue using an established precedent because it is part of their common ground with the addressee (Brennan and Clark, 1996). Under this view, speakers should misspecify more often when they are speaking to the same addressee, regardless of which partner is visible.

Dissociating visibility from participant roles required a complex setup, raising the possibility of speakers becoming confused about who could and could not hear them at any given point in time. We took several steps to ensure speakers were clear about what was going on. First, the identity of the visible partner and addressee remained fixed over a block of trials, rather than changing with each trial. Before any such block began, speakers were presented with a notification of who should be the addressee for that block. Speakers were made responsible for selecting that person as the addressee by manipulating a crossfader knob on an audio mixer. Thus, speakers had to attend to and act on the information about the addressee's identity. Second, during a warm-up phase of the experiment, speakers were given the opportunity to play the role of both addressee and non-addressee while one of the listeners took the role of the speaker. This made it clear to participants that the non-addressed partner would lack knowledge of the established referential precedents.

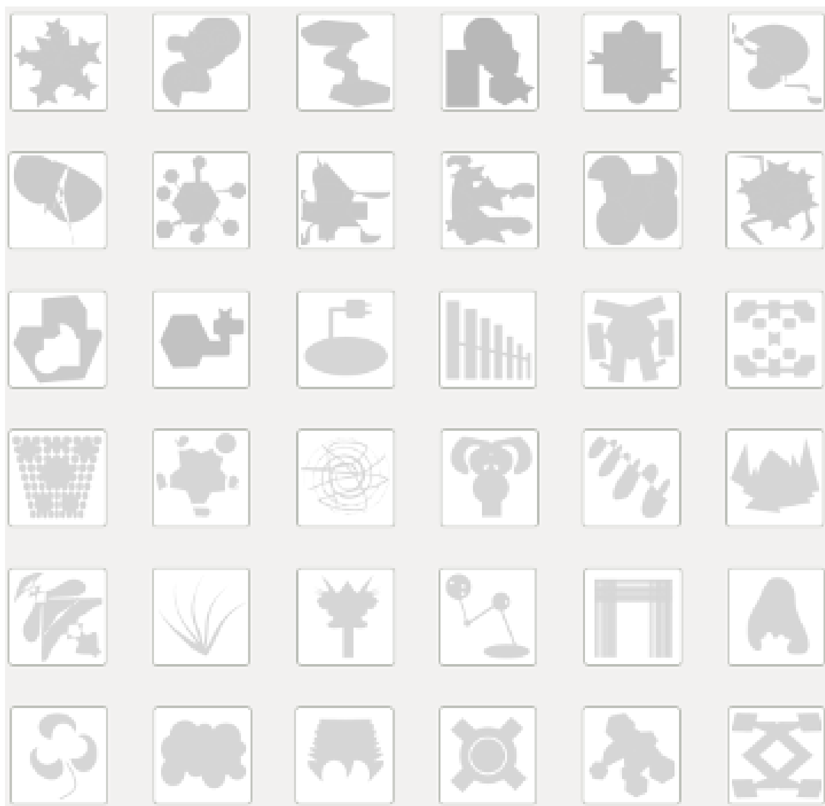

Figure 13: Unconventional targets used in Experiment 3.

Finally, we included an additional set of unconventional targets (Figure 13) which served primarily as a check on whether speakers were attending to the identity of the addressee. These unconventional targets were abstract shapes that could only be identified using structural descriptions. We included these because we have previously found that these materials elicited strong effects of common ground (Gann and Barr, 2014). Speakers gradually shorten their descriptions of unconventional objects when they refer to them repeatedly with the same addressee (Clark and Wilkes-Gibbs, 1986). Most importantly, when they describe old referents to new addressees who lack knowledge of the previous descriptions, they tend to lengthen their descriptions. To a large extent, this lengthening appears to be the result of incrementally elaborating upon the informationally-reduced descriptions arrived at with the old partner, rather than planning entirely new descriptions (Gann and Barr, 2014). Given this prior research, we expected speakers to produce longer descriptions for these targets when speaking to new addressees.

Method

A link to the pre-registration for this experiment can be found in the OSF repository.

Participants

The final dataset included data from a total of 40 University of Glasgow students (31 women and 9 men), all of whom identified themselves as native English speakers. Data from one additional participant was replaced due to continuously failing to provide informationally adequate descriptions during training (54.2% misspecifications on the last training trial before test). All participants were paid £6 or received course credit.

Design

The study had a 2x2x2 design, with factors *Shift Direction* (*Singleton-Contrast*, *Contrast-Singleton*), *Addressee* (*Same*, *Different*), and *Visible Partner* (*Same*, *Different*). The levels of all factors were administered within participants and within stimulus items.

Experimental setup and task

As in the previous experiments, speakers were to describe the target object on each trial so that an addressee could identify it. The addressee could be one of two partners (a male or female experimenter), but only one of them could actually hear the description. For the speaker, each display consisted of nine images of various objects displayed around the webcam image of the visible partner (see Figure 11). Only one of the two partners was on-screen at a given time. The partners saw only a 3x3 grid of objects, and indicated their choice by pressing a key on a number pad.

There were 48 main sets of stimuli, presented over 12 blocks of trials, with each block further subdivided into training and test phases. Each block was in one of the four conditions obtained by factorially combining the levels of *Addressee* (*Same*, *Different*) and *Visible Partner* (*Same*, *Different*), with the presentation order of the blocks determined randomly for each participant. Two of the four stimulus sets in each block appeared in the *Singleton-Contrast* condition, while the other two appeared in the *Contrast-Singleton* condition. The assignment of stimulus sets to condition was counterbalanced using eight presentation lists, with five participants randomly assigned to each list, such that each set appeared in all eight conditions of the design across participants.

Apparatus

The experimental stimuli were presented on a 19" LCD Dell desktop computer monitor (4:3 aspect ratio, resolution 1024 x 768 pixels). A microphone was placed above the participant's computer monitor to record their descriptions of the target object for each trial. Speakers controlled the crossfader on a Numark two input stereo mixer with a crossfading slider. One input had the white noise coming in on the left channel and live audio from the microphone on the right channel; the other input had the opposite configuration. The left output channel was split and fed from the mixer to one set of headphones, while the right output channel was split and fed to the other set of headphones. With this configuration, by sliding the crossfader all the way to the left, one participant would hear speech and the other would hear the white noise; sliding it all the way to the right would create the opposite situation. The two ends of the crossfader were colored so as to identify which partner would be the addressee by sliding the knob in that direction.

Video from the room with the two partners was recorded using a Logitech Pro 9000 webcam and transmitted to the speaker's display.

Materials

We re-used the stimuli from Experiment 2, except we replaced nine of the target/competitor/foil triplets with new sets, including the seven sets that were excluded from Experiment 2 due to high rates of misspecification during training. See the online repository for a list of the 48 target and competitor pairs. A third of the targets were randomly assigned to four training repetitions, a third to five training repetitions, and a third to six, forming a total of 240 training trials across all blocks for each session, and 48 test trials.

Each of the twelve blocks included two additional sets of stimuli. One of these sets consisted of filler items that we included so that a change from training to test (with a possible change of visible partner and/or addressee) did not always require a change in the description of the targets (i.e., there was no substitution of the foil/competitor from training to test). There were twelve of these sets, one for each block, half of which were constructed so that reference to the target required a modifier, and the other half so that it required no modification. The targets were repeated three times during training and once at test, forming a total of 48 trials for each session.

The other set of stimuli included in each block consisted of the unconventional targets (as described above). In each block, one unconventional target was repeated three times during training, and once at test, forming 48 additional trials for each session.

In sum, there were 384 trials in total for each experimental session: 240 training trials (4 items repeated 4-6 times in each block), 48 test trials (4 items per block), 48 fillers with targets modeled after the main stimulus items (1 item per block repeated four times), and 48 items with unconventional targets (1 item per block repeated four times).

Procedure

Upon arrival each participant was given an instruction sheet detailing the task and their role during the experiment. The instructions stated:

> You will play the role of the "Director" and will verbally name the TARGET item to one of two Matchers who will sit in a separate room from you. The figures below show the two people who will be listening to your descriptions. They will interact with you through a live webcam video. Only one Matcher will be able to hear your description at a time. The Matcher who appears on the screen may not be the person listening to your description.

The instruction sheet also contained images that provided an example of a single trial (see OSF repository for more details).

The two partners were set up in an adjoining room to the speaker and faced a single computer monitor. The partners were seated in rolling chairs, which allowed them to easily slide in front of or away from the camera, as required. The floor of the lab room was marked with tape to indicate where chairs needed to be positioned to be on or off camera. To minimize confusion for the speaker, each partner wore a colored tag that corresponded to the color of a sticker placed at each end of the crossfading slider.

Before the experiment began participants took part in a practice session that consisted of twelve training trials and four test trials. This enabled the participant to familiarize themselves with their role as speaker, as well as to experience the task from the perspective of the partner. In this manner, the participant was made aware that the partners saw the images in an entirely different spatial arrangement (Figure 11) and that only one partner at a time would be able to hear the descriptions. After practice ended, the main part of the experiment began.

On each trial, audio recording began simultaneously with the presentation of the display. The target object in each display was highlighted for the speaker by a green square. The trial ended when the addressee selected an object on the number pad. The speaker received no feedback regarding which picture the addressee selected. If the speaker failed to provide sufficient information to identify the target, the addressee would ask for clarification. Any such clarification exchanges appeared in the audio recording for the trial and were noted during later transcription.

Before each block of training trials, an on-screen notice informed the speaker which partner would appear on-screen and which partner would be the addressee. In the notice, the partners were identified by both color (the yellow and the orange partner) and first name. Partners wore different color name tags that matched the color of two stickers placed on either end of the crossfader on the audio mixer.

The on-screen notice before each training phase also indicated that the off-screen partner was to put on the blindfold. Based on the notice, the speaker slid the crossfader to the appropriate color to select the next addressee. Each of the two partners served as the training phase addressee for six of the twelve blocks. The partner who was not selected as addressee during training was always off-screen wearing a blindfold, and could only hear white noise through their headphones.

Just prior to the test phase another on-screen notice appeared indicating that the blindfold was to be removed, and designated the identity of the on-screen partner as well as the addressee. Again, the participant was responsible for sliding the crossfader to the appropriate color. So that the delay between training and test would not be confounded with condition, the notice appeared for a minimum of eighteen seconds before advancing to the next phase, which provided more than sufficient time for the partners to move into position and for the speaker to select the specified addressee.

Data Analysis

We transcribed and coded each speaker's spoken responses as described in Experiment 2. Each speaker had 96 trials for coding/transcription: the 48 test trials, and the final training trial for each stimulus item prior to the test trial. As with the previous experiments, the training trials were coded to verify that speakers were not already misspecifying the referent during training. We also transcribed each speaker's description of the twelve unconventional targets at test, and counted the number of words used in the description.

The misspecification variable was analyzed using linear mixed-effects regression, estimated with the `glmer()` function from `lme4`, with a logit link function and binomial variance. For the random effects structure, we included by-subject and by-item intercepts and also sought to include by-subject and by-item random slopes for all main effects and interactions.

Word count for unconventional targets was analyzed using linear mixed-effects regression, estimated with `glmer()` with a log link function and Poisson variance. The maximal model structure we sought to fit included by-subject and by-item random slopes for *Visible Partner*, *Addressee*, and their interaction.

Results and Discussion

We applied the same exclusion criteria for participants and stimuli in Experiment 3 as we did in Experiment 2. Based on these criteria one participant was replaced (as noted in Participants) and two of the 48 stimulus sets were removed prior to analysis, leaving 1840 total trials. Of these, an additional 131 were removed, 128 of which because the speaker did not appropriately specify the target on the final training trial, and three because the speech could not be determined due to poor recording quality.

Misspecification rate

The logistic regression model of misspecification rate did not converge with maximal random effects. We fit a second model in which we reduced the random effects structure until convergence was reached and no singularity message was

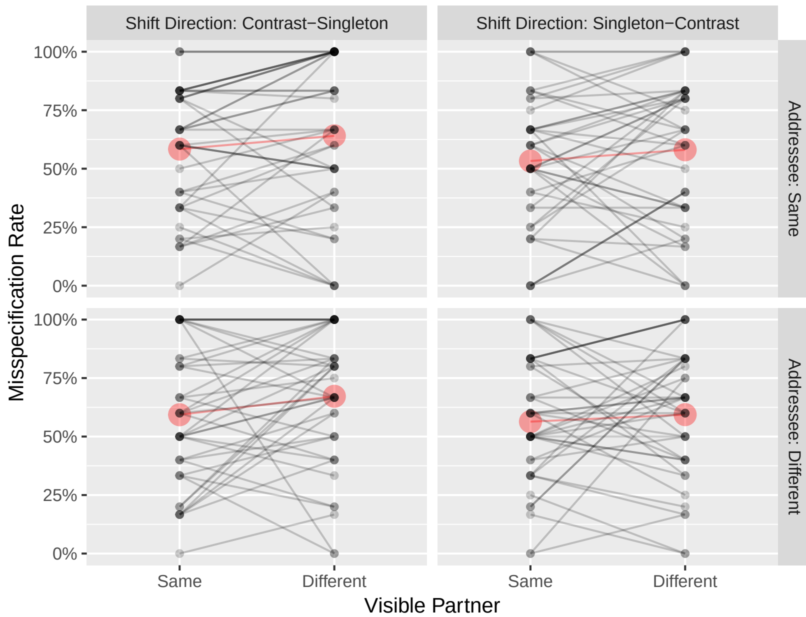

Figure 14: Misspecification rate by Visible Partner, Addressee, and Shift Direction. Connected black points are individual participants and red points are grand means.

encountered. The reduced model contained by-subject random intercepts, by-subject random slopes for *Visible Partner*, the *Visible Partner-by-Addressee* interaction, and the three way interaction, with covariances constrained to zero; by-item random intercepts, by-item random slopes for *Shift Direction*, *Shift Direction-by-Visible Partner* interaction, the three way interaction, and with covariances also constrained to zero.

The key prediction concerned whether the misspecification rate was higher when the visible partner was the same at test as at training. There was little evidence to support this prediction (Figure 14). Misspecifications were observed on 57.7% of trials where the visible partner was the training partner, compared to 63.0% where the visible partner was the other partner. This difference was not significant (pre-registered one-tailed test), $\beta = -0.32$, $SE = 0.12$, Wald $z = -2.66$, $p = 0.996$, and was in fact showing a numerical trend in the opposite direction from what was predicted. No other effects were significant. (Because

Table 4: Distribution of utterance types in the Contrast-Singleton condition, Experiment 3. NO = no misspecification; PR = pre-nominal modifier; PO = post-nominal modifier; DE = deleted adjective; AS = Addition by self-repair; AO = addition by other-repair. Categories PR, PO, DE, AS, and AR were counted as misspecifications.

Visible Partner	Addressee	NO	PR	PO	DE	AS	AO
Same	Same	41.2%	48.9%	9.0%	0.0%	0.9%	0.0%
Same	Different	41.4%	45.5%	8.6%	3.2%	1.4%	0.0%
Different	Same	34.1%	50.9%	11.4%	2.7%	0.9%	0.0%
Different	Different	33.2%	51.4%	13.6%	1.8%	0.0%	0.0%

Table 5: Distribution of utterance types in Singleton-Contrast condition, Experiment 3. NO = no misspecification; PR = pre-nominal modifier; PO = post-nominal modifier; DE = deleted adjective; AS = Addition by self-repair; AO = addition by other-repair. Categories NO, DE, AS, and AO were counted as misspecifications.

Visible Partner	Addressee	NO	PR	PO	DE	AS	AO
Same	Same	2.0%	28.2%	15.8%	0.0%	21.8%	32.2%
Same	Different	3.3%	23.0%	19.6%	0.5%	20.6%	33.0%
Different	Same	2.5%	22.5%	17.5%	0.0%	18.0%	39.5%
Different	Different	2.4%	23.2%	17.5%	0.0%	18.5%	38.4%

this experiment had three factors and thus more parameters than the previous experiments, we present all remaining parameter estimates in Table 6 instead of in the main text.)

Table 6: Parameter estimates, standard errors, test statistics and p-values for analysis of misspecification rate (see main text for Visible Partner (VP) results, which was a pre-registered one-tailed test).

effect	beta	SE	Wald z	p
Intercept	0.56	0.18	3.09	0.002
Shift Direction (SD)	-0.18	0.17	-1.07	0.287
Addressee (A)	-0.04	0.11	-0.33	0.739
SD:VP	0.17	0.23	0.74	0.461
SD:A	0.12	0.22	0.53	0.593
VP:A	0.02	0.24	0.07	0.941
SD:VP:A	-0.13	0.50	-0.25	0.801

There was little evidence for the prediction that misspecification would be higher when the training partner was the partner who was visible at test. One question is whether participants were sensitive at all to the visible partner manipulation; perhaps speakers were sensitive, but used the information in different ways—some subjects showing greater misspecification when the visible partner matched, others showing greater misspecification when it mismatched—with these individual differences yielding no overall systematic effect. To test this, we first performed an significance test of the by-subject random slope for Visible Partner (this test was not part of pre-registration plan). The estimated by-subject random slope of 0.35 was not significantly different from zero, $\chi^2(1) = 1.24$, $p = 0.265$. (We did not test the significance of the by-item random slope because the model estimated this parameter to be zero.)

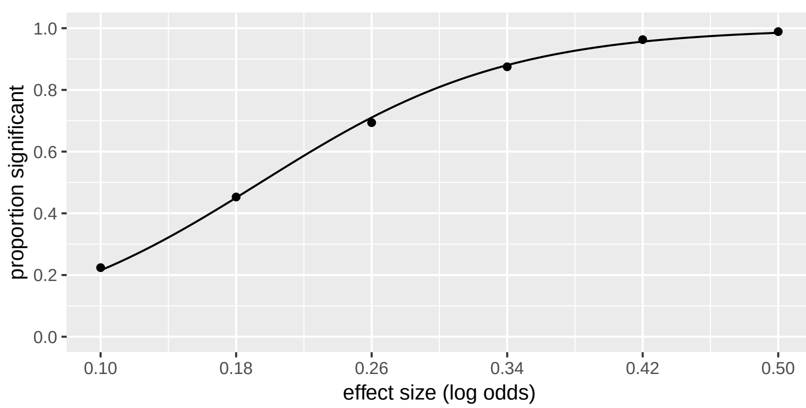

Figure 15: Sensitivity analysis for Experiment 3. The points are the proportion of significant tests at each of the six effect sizes examined; the line is a fit from a logistic regression model to allow smooth interpolation.

We also performed a sensitivity analysis for the test of the main effect of *Visible Partner*, following the same procedure described for Experiment 1, with the difference that the effect size range was from .1 to .5 logits. (This analysis was also not pre-registered.) Results are shown in Figure 15. The sensitivity analysis suggests the test of *Visible Partner* had: (1) 80% power to detect an effect of at least 0.3 logits, corresponding to a difference in misspecification rate of about 5.1% between conditions (e.g., 58.6% in the different partner versus 63.7% in the same partner condition); (2) 90% power to detect an effect of at least 0.35 logits, or a difference in response rate of 6.7% (e.g., 57.5% versus 64.2%); and (3) 95% power to detect an effect of at least 0.41 logits, or a difference in response rate of 6.5% (e.g., 58.1% versus 64.6%).

Word count for unconventional targets

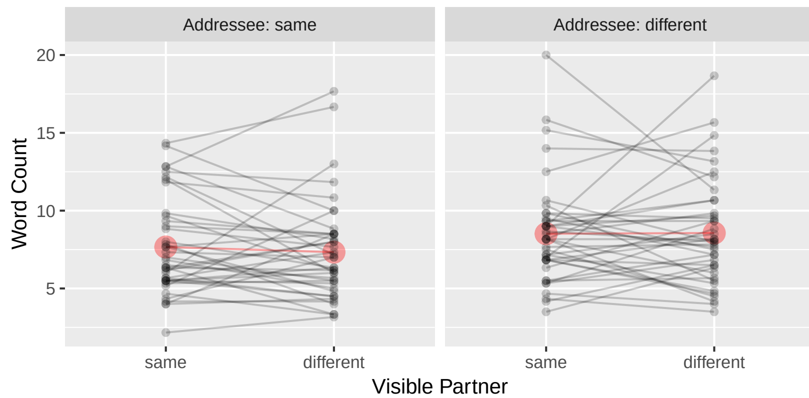

Figure 16: Word count for unconventional referent descriptions by Visible Partner and Addressee. Connected black points are individual participants and red points are grand means.

For the model of word count, we fit a generalized linear mixed-effects model with Poisson distribution function and log link. The maximal model returned a singularity message, and so we fit a reduced model with a by-subject random intercepts, by-subject random slopes for *Visible Partner*, *Addressee*, and the *Visible Partner-by-Addressee* interaction, with covariances constrained to zero; by-item random intercepts, by-item random slopes for *Addressee*, and covariances constrained to zero. The data are shown in Figure 16.

There was a significant effect of *Addressee* on word count, with longer descriptions given when the partner at test was different from the training partner, $M = 8.5$ ($SD = 5.5$), compared to descriptions given to the same partner, $M = 7.5$ ($SD = 4.7$), $\beta = -0.26$, $SE = 0.06$, Wald $z = -4.48$, $p < .001$.

There was little evidence for any effect of *Visible Partner* on word count; for descriptions in the *Different Partner* condition, $M = 7.9$ ($SD = 5.4$), versus descriptions in the *Same Partner* condition, $M = 8.1$ ($SD = 4.9$), $\beta = -0.01$, $SE = 0.03$, Wald $z = -0.30$, $p = 0.762$.

Finally, there was little evidence for a *Visible Partner-by-Addressee* interaction. When the visible partner was the same as in training, speakers produced descriptions that were on average 0.9 words longer for the new addressee; this effect was not significantly different from the case where the visible partner was different from the training partner, where speakers produced descriptions that were on average 1.2 words longer for the new addressee, $\beta = 0.01$, $SE = 0.08$, Wald $z = 0.19$, $p = 0.850$.

General Discussion

Over three experiments, we sought to test whether ordinary memory processes—as embodied in the encoding specificity principle (Tulving and Thomson, 1973)—influence the selection of information in the generation of referential descriptions. The basic logic of the experiments was to have speakers entrain on particular descriptions for referents in specific contexts, and then present the same referents in a context with different informational requirements but varying in similarity to the training context. The key prediction was that speakers' tendency to use the (no longer appropriate) entrained-upon description would vary as a function of the similarity between the training and test contexts. Support for this prediction was weak and inconsistent: no statistically reliable effect in Experiment 1, with a difference in means of less than 1%; a statistically significant congruency effect in Experiment 2 (pre-registered one-tailed test, p =0.030), but with a difference of less than 5%; and finally, a numerical difference of 5% in Experiment 3, but in the wrong direction. All experiments were pre-registered and attained 80% power to reject a raw effect of around .4 logits, corresponding to a difference in misspecification rate of about 5%.

Despite limited overall support for the main prediction that training-test similarity would modulate misspecification rates, all three experiments show strong memory effects, inasmuch as speakers consistently misspecified targets across all three experiments: 16% in Experiment 1, 82% in Experiment 2, and 60% in Experiment 3. These overall rates indicate that speakers did retain information from training episodes, since the specific misspecifications that took place (e.g., calling a candle 'the unmelted candle') would be extremely unlikely to occur in the absence of the training experiences; moreover, the eye data from

Experiment 2 indicated strong and detailed memory for the training display configurations, although the impact on speech was limited. Finally, these mostly null effects of detailed memory representations are contrasted with positive evidence for a common ground effect in Experiment 3, where speakers lengthened descriptions of unconventional referents for new addressees. Taken together, these findings support the idea that speakers do maintain detailed representations about past referring episodes, but these representations have little role in the message generation process, even when the representations are related to the identity of an interacting conversational partner. Instead, it appears that much of message generation is driven by coarse-grained memory representations that do not contain much more information than the identity of the target referent and the label given to it on previous occasions.

It is illustrative to consider these findings in relation to recent findings from comprehension. Episodic effects on comprehension have been studied in a similar paradigm, in terms of whether reference resolution is facilitated when listeners hear expressions repeated in the voice of the speaker who established the precedent. Although early experiments failed to find such facilitation (Barr and Keysar, 2002; Metzing and Brennan, 2003), it was eventually detected in later experiments that used more sensitive measures and larger samples (Brown-Schmidt, 2009). A meta-analysis suggested these effects are likely to exist, but are small and fleeting (Kronmüller and Barr, 2015) especially when compared to the very large and reliable partner-independent effects. In short, abstract symbolic memory representations, such as the association between a referent and a referring expression, appear to have strong impacts on language processing, but the role of more detailed episodic representations appears marginal at best. That said, our findings for production are best viewed as limiting the explanatory scope of ordinary memory models, rather than as an overall rejection of this view. Our studies have only looked only at short-term memories formed within the confines of the laboratory, and perhaps repetition across a longer time frame could produce larger effects.

Another consideration is that across all experiments, we used experimenters as listeners rather than actual participants. One possibility is that because our experimenters were practiced at the task, back-and-forth interaction was more limited than it would be with uninformed listeners, and perhaps speakers attended less to the referring context than they would otherwise, thus forming impoverished representations. Against this interpretation, we note that we did find strong partner effects in Experiment 3 with the unconventional targets, which demonstrates that speakers were treating the two listeners as having different knowledge and did in fact encode information about the context. Furthermore, it could be argued that using real listeners could lead to weaker encoding of context, since they would be likely to produce more variable responses, respond with greater delay, and their relative unfamiliarity and uncertainty could distract attention from the displays onto the interaction itself.

Our results do not entirely reject the claim of the ordinary memory view that memory processes can serve as a proxy for common ground in message generation; rather, they help set boundary conditions for this claim. As already noted, speakers' memory representations were only established through repetition within the context of a single conversation with a single partner, and so our findings may not generalize to representations developed through repetition across multiple conversations taking place over longer time intervals. Still, retaining and using information from the current conversation is a component of common ground, embodied in the *linguistic co-presence* heuristic cited in the Introduction. Our results therefore suggest it is unlikely that ordinary memory processes can effectively substitute for this heuristic, but they leave the *physical co-presence* and *community membership* heuristics untouched. Interestingly, even using an ecologically valid memory cue—the image of the training partner (Experiment 3)—did not appear to strengthen effects relative to the more abstract cues used in the first two experiments.

To avoid a possible misunderstanding of our position, we do not intend to claim that ordinary memory processes are the sole determinant of information availability during message generation. There are likely to be many perceptual or social-cognitive factors operating outside of these processes that are worthy of further study. Indeed, our Experiment 3 provided evidence that speakers used common ground, lengthening their descriptions of old referents to provide more information when speaking to a partner with whom previous descriptions had not been shared, replicating both Gann and Barr (2014) and Ahn and Brown-Schmidt (2020). Also like Gann and Barr (2014), we found that speakers showed partner effects when describing abstract referents but not when using modifiers. What might explain this apparent conflict? Whereas modified descriptions such as “the big W” or “the unmelted candle” are likely to be mostly pre-planned, longer descriptions that speakers produce for abstract objects (e.g., “it looks like three snail-shapes stacked on top of each other”) afford opportunities for incremental construction, such that their content may be shaped by self- and other-monitoring processes. Supporting this view, Gann and Barr (2014) observed that despite producing longer descriptions, speakers’ onset latencies were no greater when they spoke to new addressees, which indicates that the extra content was not part of the original plan. Also, there was evidence that the length of speakers’ utterances could be predicted by hesitation behaviors emitted by the addressee, supporting the idea that the extra content resulted from monitoring of the partner.

The question of how speakers select information in language production remains one of the least studied, and thus, most mysterious aspects of language production. One point that scholars can agree on is that much of what speakers choose to say seems to be driven in large part by information availability, but the concept of ‘availability’ remains a poor explanatory construct. While ordinary memory processes are inevitably involved, what they deliver up to production processes are largely abstract symbolic representations, which makes it unlikely that these processes serve as an effective proxy for common ground in everyday conversation.

References

Ahn, S., Brown-Schmidt, S., 2020. Retrieval processes and audience design. Journal of Memory and Language 115, 104149.

Barr, D. J., Keysar, B., 2002. Anchoring Comprehension in Linguistic Precedents. Journal of Memory and Language 46, 391–418.

Barr, D. J., Levy, R., Scheepers, C., Tily, H. J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language 68, 255–278.

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015a. Parsimonious mixed models. arXiv preprint arXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015b. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48.

Brennan, S. E., Clark, H. H., 1996. Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, & Cognition 22, 1482–1493.

Brown-Schmidt, S., 2009. Partner-specific interpretation of maintained referential precedents during interactive dialog. Journal of Memory and Language 61, 171–190.

Brown-Schmidt, S., Horton, W. S., 2014. The influence of partner-specific memory associations on picture naming: A failure to replicate Horton (2007). PloS one 9, e109035.

Chun, M. M., Jiang, Y., 1998. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive psychology 36, 28–71.

Clark, H. H., 1973. The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. Journal of Verbal Learning and Verbal Behavior 12, 335–59.

Clark, H. H., Marshall, C. R., 1981. Definite reference and mutual knowledge. In: Joshe, A. K., Webber, B. L., Sag, I. A. (Eds.), Elements of Discourse Understanding. Cambridge University Press, Cambridge, pp. 10–61.

Clark, H. H., Murphy, G. L., 1982. Audience design in meaning and reference. In: Le Ny, J., Kintsch, W. (Eds.), Language and comprehension. North Holland Publishing, Amsterdam, pp. 287–299.

Clark, H. H., Wilkes-Gibbs, D., 1986. Referring as a collaborative process. Cognition 22, 1–39.

Dell, G. S., Brown, P. M., 1991. Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener". Erlbaum, Hillsdale, N.J.

Engelhardt, P. E., Demiral, B., Ferreira, F., 2011. Over-specified referring expressions impair comprehension: An ERP study. Brain and Cognition 77, 304–314.

Ferreira, V. S., Dell, G. S., 2000. Effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive Psychology 40, 296–340.

Fukumura, K., van Gompel, R. P. G., 2012. Producing Pronouns and Definite Noun Phrases: Do Speakers Use the Addressee's Discourse Model? Cognitive Science 36, 1289–1311.

Fussell, S. R., Krauss, R. M., 1992. Coordination of knowledge in communication: Effects of speakers' assumptions about others' knowledge. Journal of Personality and Social Psychology 62, 378–391.

Gann, T. M., Barr, D. J., 2014. Speaking from experience: Audience design as expert performance. Language, Cognition and Neuroscience 29, 744–760.

Grice, H. P., 1975. Logic and conversation. In: Cole, P., Morgan, J. (Eds.), Syntax and Semantics 3: Speech acts. Academic Press, New York, pp. 41–58.

Grodner, D., Sedivy, J. C., 2011. The effect of speaker-specific information on pragmatic inferences. MIT Press, Cambridge, MA, pp. 239–271.

Hintzman, D. L., 1986. "Schema abstraction" in a multiple-trace memory model. Psychological Review 93, 411–428.

Horton, W. S., 2007. The influence of partner-specific memory associations on language production: Evidence from picture naming. Language and Cognitive Processes 22, 1114–1139.

Horton, W. S., Gerrig, R. J., 2005. The impact of memory demands on audience design during language production. Cognition 96, 127–142.

Horton, W. S., Gerrig, R. J., 2016. Revisiting the memory-based processing approach to common ground. Topics in Cognitive Science 8, 780–795.

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

Kahneman, D., Frederick, S., 2002. Representativeness revisited: Attribute substitution in intuitive judgment. Cambridge University Press, Cambridge, UK, pp. 49–81.

Kronmüller, E., Barr, D. J., 2015. Referential precedents in spoken language comprehension: a review and meta-analysis. Journal of Memory and Language 83, 1–19.

Logan, G. D., 1988. Toward an instance theory of automatization. Psychological Review 95, 492–527.

Logan, G. D., Etherton, J. L., 1994. What is learned during automatization? The role of attention in constructing an instance. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 1022–1050.

Mayo, D. G., 2018. Statistical inference as severe testing. Cambridge: Cambridge University Press.

Metzing, C., Brennan, S. E., 2003. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. Journal of Memory and Language 49, 201–213.

Nadig, A. S., Sedivy, J. C., 2002. Evidence of perspective-taking constraints on children's on-line reference resolution. Psychological Science 13, 329–336.

Pickering, M. J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 1–22.

Posner, M. I., Keele, S. W., 1968. On the genesis of abstract ideas. Journal of Experimental Psychology 77, 353–363.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., Boyes-Braem, P., 1976. Basic objects in natural categories. Cognitive Psychology 8, 382–439.

Rossnagel, C., 2000. Cognitive load and perspective taking: Applying the automatic-controlled distinction to verbal communication. European Journal of Social Psychology 30, 429–445.

Ryan, J. D., Althoff, R. R., Whitlow, S., Cohen, N. J., 2000. Amnesia is a deficit in relational memory. Psychological Science 11, 454–461.

Simpson, I. C., Mousikou, P., Montoya, J. M., Defior, S., 2013. A letter visual-similarity matrix for latin-based alphabets. Behavior Research Methods 45, 431–439.

Tulving, E., Schacter, D. L., 1990. Priming and human memory systems. Science 247, 301–306.

Tulving, E., Thomson, D. M., 1973. Encoding specificity and retrieval processes in episodic memory. Psychological Review 80, 352–373.

Wardlow Lane, L., Groisman, M., Ferreira, V. S., 2006. Don't think of a pink elephant! Speakers' control over leaking private information during language production. Psychological Science 17, 273–277.