Running head: SPEAKING FROM EXPERIENCE

Speaking from experience:
Audience design as expert performance

Timothy M. Gann

University of California, Riverside

Dale J. Barr

University of Glasgow

Abstract

When speakers generate referential descriptions, they take their addressees' needs into account through processes of *audience design*. In this paper we consider audience design as a kind of expert performance, in which skilled behavior is the result of an interplay between memory and attention. We suggest that attending to referents with a referential goal in mind results in the obligatory retrieval of previous descriptions, and that these previous descriptions are adapted to current purposes via monitoring and adjustment. In an experiment, speakers gained experience describing certain referents with a given addressee, and then later described these referents to either the same or a different addressee. When describing old referents, speakers relied on remembered descriptions, adding detail via monitoring and adjustment to meet an addressee's less informed perspective. However ~~In contrast,~~ speakers were unable to elide information that was no longer relevant, resulting in a high rate of referential overspecification. These findings provide new insights into partner-adaptation, and highlight the value of theories of expertise for the study of audience design in language production.

Speaking from experience:

Audience design as expert performance

In conversation, speakers face a recurring question: *What shall I call this?* Almost anything that a speaker might wish to say requires mentioning some referent (i.e., object, event, or entity). The efficiency and apparent ease with which speakers produce referential expressions masks an underlying cognitive complexity arising from the multiple ways in which any given thing can be conceptualized and spoken about. From among these possibilities, speakers must choose an expression that meets the informational needs of their interlocutors. This process of adapting one's speech to another's needs is known as *recipient design* (Sacks, Schegloff, & Jefferson, 1974) or *audience design* (Clark & Murphy, 1982).

On one account, a critical kind of information underlying successful audience design of referring expressions is *common ground*, information that interlocutors have good reason to believe is mutually shared (Clark & Marshall, 1981). Speakers assess common ground on the basis of several *copresence heuristics*, which includes physical co-presence (whether a speaker and an addressee can both perceive an entity), linguistic co-presence (whether information forms part of their discourse record), and community knowledge (the knowledge and beliefs that are shared among the communities to which the speaker and addressee belong). A key psycholinguistic question is how considerations of common ground influence speakers' choices when producing an utterance.

Studies of audience design in reference generally find that speakers do not always optimize expressions to listeners' informational needs (Engelhardt, Bailey, & Ferreira, 2006; Ferreira & Dell, 2000; Horton & Keysar, 1996; Wardlow Lane & Ferreira, 2008; Wardlow Lane, Groisman, & Ferreira, 2006; Wardlow Lane & Liersch, in press). Evidently, speakers' tendency to include information in their descriptions depends not only on the accessibility of information to the addressee, but also on its accessibility to the speaker's

own production processes (Wardlow Lane & Ferreira, 2008; Wardlow Lane et al., 2006; Wardlow Lane & Liersch, in press). This dependence on speaker-internal factors would seem to imply that speakers would tend to regularly include information in descriptions that is misleading or unhelpful for the listener. However, this tendency is mitigated by at least two factors. *Monitoring-and-adjustment models* assume that a secondary "self-monitoring" stage of language production uses common ground (and/or monitoring of the addressee for evidence of understanding) to adapt an original utterance plan to meet listeners' needs; this secondary stage, due to its effortful nature, is only deployed if speakers have sufficient time and cognitive resources (Brown & Dell, 1987; Horton & Keysar, 1996). A second mitigating factor is that memory operates in a context-sensitive manner through the encoding specificity principle (Tulving & Thomson, 1973), such that information is retreived as an increasing function of the similarity of the retrieval context to the encoding context. *Memory-based models* assume that this encoding specificity allows partner-oriented speech to simply "emerge" from memory, because addressees serve as retrieval cues that contextualize the memory retrieval process, increasing the likelihood of retrieving shared information (Gerrig & McKoon, 1998; Horton & Gerrig, 2002).

Although much has been learned about the processes involved in audience design, the field is still far from a comprehensive theory. One reason is that processing and representational issues are generally treated separately. Monitoring-and-adjustment models focus strictly on processing issues, and take the existence of appropriately structured representations for granted. In contrast, memory-based models tend to focus on representational issues, but do not consider how those representations are themselves structured as a result of previous processing in the domain.

One way to unify the various theoretical approaches to audience design is to cast them into the broader theoretical framework of expert performance. Theories of expert performance view skilled behavior as the result of extensive experience in a domain,

through which an expert acquires a large database of memory representations that map problem instances to solutions, greatly facilitating problem solving (Anderson, 1996; Chase & Simon, 1973; Logan, 1988; Simon, 1996). One theoretical framework that seems particularly applicable is Logan's theory of skill acquisition, the Instance Theory of Automaticity (ITA) (Logan, 1988). According to the ITA, on one's first encounter with a problem, the problem is solved algorithmically, in a manner that draws upon high-level reasoning and problem solving processes. This "processing episode" is stored in memory. Over repeated encounters with the problem, memory traces build up that directly map the problem stimulus to the solution. Upon each encounter with the problem, the memory retrieval route "races" with algorithmic processes toward a solution, with the retrieval route predominating in skilled performance. An important assumption behind ITA is that memory retrieval is an obligatory consequence of attention; that is, one cannot deliberately "preempt" retrieval processes from operating even in contexts where one has foreknowledge that former solutions will no longer apply.

The ITA framework has been applied to language processing previously, in the work of Rawson (2010) on reading comprehension. Here, we extend the ITA to audience design issues in referential description in the following manner. One's first encounter with a referent is likely to involve all manner of reasoning of a Gricean type, as one searches for a way to linguistically categorize a referent that distinguishes it from viable alternatives in the discourse context. The resulting description becomes associated with the cognitive antecedent conditions that instantiated the process (i.e., attending to the referent with the intent to produce a definite description), and this "processing episode" is stored in memory. Reinstantiating the same antecedent conditions elicits, in an obligatory manner, the retrieval of previous descriptions.

Although an expert performance framework such as Logan's has great promise, we still lack critical knowledge for how such a model should be specified. What constitutes a

"processing episode" in relation to audience design? What is the nature of the memory traces left by the production process? How abstract are they, and how are they molded by attention? Do memory and algorithmic routes operate in parallel, or are algorithmic processes only called upon when the memory route fails to deliver up a sufficiently strong signal? In what ways can attention modulate retrieval processes?

In this paper, we focus mainly on issues related to the interplay between memory and attention, leaving questions about the nature of processing episodes for future research. We report results from a language production experiment exploring the conditions that lead speakers to "misspecify" referents—that is, to provide listeners with either more or less information than would be optimal for their understanding. The basic logic was to give speakers the opportunity to develop memory routines for describing particular referents, and then to test whether they continued to rely on these routines when the context changed in critical ways. This approach helps to distinguish components of the production of referring expressions that operate relatively automatically (in the sense that they rely on memory retrieval) from those that are more algorithmic (delibrative, effortful) in nature.

Explaining referential misspecification

Referring to an object in conversation is a cooperative act; and, as such, speakers should provide addressees with no more and no less information than they would need to identify the referent (Grice, 1975). However, under certain conditions, speakers systematically fail to meet this requirement. The most commonly observed finding is overspecification, with speakers using modifiers when relevant contextual support is absent (e.g., calling a circle *the black circle* when it is the only circle in the array). Pechmann (1989), using eyetracking, found evidence that overspecification reflects incremental

processing: speakers crafted an utterance based on salient properties of the target and begin articulating it before having fully checked the context for its adequacy (Pechmann, 1989). Overspecification occurs not only when contextual support is wholly unavailable to the speaker (Deutsch & Pechmann, 1982; Engelhardt et al., 2006; Pechmann, 1989) but also when it is available to the speaker but not to the addressee (Horton & Keysar, 1996; Nadig & Sedivy, 2002; Wardlow Lane et al., 2006; Wardlow Lane & Ferreira, 2008). Brennan and Clark (1996) found that speakers who had entrained on a subordinate level term to distinguish an object from another of the same category (e.g., calling a shoe a *loafer* to distinguish it from a sneaker) continued to use these specific terms even when the contextual support (e.g., the other shoe) was removed from the array.

Speakers seem much less likely to underspecify than to overspecify referents (Ferreira, Slevc, & Rogers, 2005). Although young children routinely underspecify referents, this is rarely observed in adults (Deutsch & Pechmann, 1982). However, studies with adult participants have not yet created circumstances that would be likely to elicit underspecifications. From an expertise point of view, a speaker S would seem most likely to underspecify referent R under the following three conditions: (1) S has routinized a description for R in context C_1; (2) the current referring context C_2 requires more information about R than C_1; (3) during production, S has no "online" feedback from an addressee that might inform S that more information was needed. One study that came close to satisfying these conditions was Wilkes-Gibbs and Clark (1992) (see also Galati & Brennan, 2010 for a test of underspecification in the context of narrative descriptions). In Wilkes-Gibbs and Clark (1992), speakers described abstract figures over a number of turns to a particular addressee, and their descriptions became shorter with each reference turn (Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964). Following this, speakers had to describe the same figures again to a different addressee, whose knowledge of the previous descriptions varied. When the new addressee was a naïve participant, speakers

gave longer descriptions than when he or she had been a "side participant" to the original conversation (i.e., a silent participant). However, neither this study nor the study by Galati and Brennan (2010) satisfied the third condition, since addressees could freely give speakers feedback, possibly preempting any underspecification. With such prompting, addressees might have turned what would have otherwise been an underspecified description into an appropriately specified one.

Audience design is a complex process that takes place in complex environments. Consequently, a wide variety of explanations can be given for why a speaker overspecifies (or avoids underspecifying) a given referent in a given context. This introduces a number of practical challenges to acquiring the critical evidence needed to elaborate a mechanistic account of audience design. We identify these challenges below, and discuss how we addressed them through our experimental design and procedure. They are: (1) assessing what counts as "optimal" information for identifying a given referent in a given context; (2) distinguishing adaptations that speakers produce because of changes in salience of information to themselves, as opposed to changes in the common ground; (3) distinguishing individual-level from interactional-level explanations of utterance content; and (4) distinguishing adaptations that speakers achieved during initial utterance planning from those achieved during later monitoring stages.

The first challenge is determining what counts as an "optimal" amount of information for a given addressee in a given context. A speaker may call a referent the "red circle" even when there is no circle of a contrasting color in the referential array, but it is possible that a generic listener would nonetheless find this "extra" information useful in identifying the referent (Arts, Maes, Noordman, & Jansen, 2011; Mangold & Pobel, 1988). Thus, it may be dangerous to make assumptions about what is optimal from a purely logical (i.e., Gricean) approach. Under the assumption that members of a language community are likely to have highly similar cognitive representations (see Barr, 2004;

Garrod & Doherty, 1994; Steels, 1997), one good estimate of what information might be optimal for a generic addressee is the information that speakers provide on their first description of a referent. The question then is whether speakers, after being given experience describing particular referents, would give addressees more or less information than this baseline amount compared to a speaker describing the same referent in the same context, but who lacks that experience. Thus, a primary manipulation in the experiment was *Novelty*: whether or not a given referent was new or old for the speaker.

Speakers who have gained experience describing referent R in context C_1 may adapt their descriptions when they later describe R again in new context C_2. This context-sensitivity would seem to be *prima facie* evidence for audience design, as the change in context effects a change in the common ground; however, it also changes the information available to the speaker. Any rigorous test of audience design must hold speakers' knowledge constant while manipulating their beliefs about addressees' informational needs (Keysar, 1997). One standard way of doing this is to use multiple addressees in the experiment, one of whom shares the speaker's experience, and the other of whom does not (see also Brennan & Clark, 1996; Horton & Gerrig, 2002; Galati & Brennan, 2010; Wilkes-Gibbs & Clark, 1992). To this end, the second manipulation in our experiment was *Addressee*, with half of the speakers speaking to the same partner throughout the experiment, and the other half participating with two addressees, one of whom shared the speaker's experience, and the other of whom did not.

The processes that result in the output of an utterance are generally thought of as pertaining only to the speaking interlocutor. However, it is clear that addressees can play an active role in determining an utterance's informational content through cues they provide to the speaker (Bavelas, Coates, & Johnson, 2000; Clark & Wilkes-Gibbs, 1986; Kraut, Lewis, & Swezey, 1982; Schober & Clark, 1989), and that speakers monitor these cues while speaking (Clark & Krych, 2004). This raises a third challenge in studying

audience design, which is distinguishing adaptations made on the speaker's own initiative, without prompting from addressee feedback (*self-prompted adaptations*) from adaptations speakers make in response to feedback from the listener (*other-prompted adaptations*) (Barr & Keysar, 2006). For instance, Brown and Dell (1987) found that speakers were more likely to mention atypical than typical instruments used to perform an action (e.g., stabbing someone with an icepick versus a knife). This tendency appeared to be driven by the speaker's own knowledge rather than by the speaker's perception of the addressees' needs. However, speakers in their study were speaking to practiced confederate listeners. Lockridge and Brennan (2002) found that when speakers spoke to naïve addressees with real informational needs, they were more likely to adapt their speech for the listener. These divergent findings suggest that certain adaptations might not be produced in the absence of feedback from the addressee. Speakers are sensitive not only to the presence of certain cues or backchannels that evince understanding, but also to the absence of expected cues. This means that in any given case where there is even a *potential* for feedback, it many not be possible in principle to distinguish between self- and other-prompted adaptations. But the situation is not as hopeless as it may seem: when the potential for feedback is wholly eliminated, and speakers have knowledge of this before starting to plan their descriptions, any adaptations they produce could only be self-generated. Thus, *Feedback Availability* was an additional factor in the experiment; speakers described referents with foreknowledge of whether feedback from the addressee would be available or unavailable.

Finally, to the extent that speakers spontaneously design messages that reflect their audience's needs, it is possible to further distinguish whether they do during initial utterance planning (which, following Horton & Keysar, 1996, we refer to as *initial design*), or whether they do so through monitoring-and-adjustment, after they have already begun articulating the utterance. One way to distinguish between these possibilities is through

the use of online measures. To this end, we tracked speakers' eye movements (Pechmann, 1989) and measured their speech onset latency. If a speaker gives a longer description of an old referent to a new partner than to an old partner, but without incurring additional planning time, that would imply that the adaptation was mediated by monitoring-and-adjustment mechanisms and not initial design. Furthermore, we can use speakers' eye movements to examine their use of common ground during planning and monitoring and adjustment. Speakers' common ground with the addressee will differ depending on whether they share knowledge of their prior descriptions with the addressee. When they do, listeners can disambiguate the reference using the form of the expression itself (e.g., "the pac-man shape" implies "the referent we have called 'the pac-man shape' on past occasions", not just any pac-man-like shape). This makes the interpretation of such expressions less dependent on the actual set of alternatives in the visual array. Therefore, one would expect less scanning of the array than in the case when the referent is old for the addressee than when it is new, since in the latter case the relevant common ground is the set of physically co-present alternatives. To the extent there is greater scanning with a new partner prior to speech onset gives evidence for the use of common ground in initial design.

The present study

To summarize, speakers gained experience describing a set of referents during a "training" phase and then described these referents again during a "test" phase where the context was either the same or different from the training phase. We manipulated three factors in our experiment: *Novelty*, *Addressee*, and *Feedback Availability*, and measured speakers' verbal descriptions as well as online processing (speech onset latency and visual scanning). Let us now describe the experimental task and design in more detail.

Participants played the role of speaker in a referential communication game, half of

them playing with one additional participant who was given the addressee role, and the other half playing with two additional participants who alternated turns as addressee. The speaker and addressee sat side by side facing a single computer monitor. In each trial of the game, the speaker saw five pictures displayed at the corners of an imaginary pentagon, and was (privately) informed of the location of a "target" object. The speaker's task was to describe the object without indicating its location so that the addressee could identify it (and select it using a computer mouse). Speakers could observe the movement of the mouse cursor, and thus had an online indicator of how effective their expressions were.

The experiment was divided into a series of blocks, each of which consisted of a "training" phase followed by a "test" phase. In the training phase, speakers developed expertise in referring to various objects. They referred to each object that would later appear in the test phase a total of five times. In the test phase, they referred to these old referents as well as some new referents (that were neither seen nor referred to during training). During the training phase, addressees were free to provide feedback to the speaker.

Speakers referred to two different kinds of referents, *unconventional* and *conventional* targets, in order to create opportunities for under- and over-specification, respectively. The unconventional targets were unusual, abstract figures (see Figure 1). They were included because speakers would lack any experience with these objects, and would have to come up with new descriptions, which they could abbreviate over time. The key question was whether in the test trials, speakers would use abbreviated descriptions at the same rate with a new addressee (who lacked the speaker's expertise) as with the old, and if not, whether the expanded descriptions given to new listeners were self- or other-prompted.

Conventional targets were everyday objects (e.g., candles, shoes, etc.) and were included to induce referential overspecification. During training, each conventional target

appeared in the display alongside another "competitor" object of the same category. For example, a candle appeared alongside another melted version of the candle, requiring it to be called, e.g., *the unmelted candle*. The target was always the more prototypical member of the category, such that mentioning the distinguishing feature would be unusual in the absence of contextual support. In test trials, the competitor was absent from the display, such that using the modified description (e.g., *unmelted candle*) would constitute overspecification. The question was whether the overspecification rate would vary with the conversational partner, as well as with the availability of feedback.

Method

Participants

The experiment involved 80 participants (47 females and 33 males), who formed 16 triads and 16 dyads, drawn from the undergraduate population of the University of California, Riverside. The experimenter randomly assigned members of each dyad or triad to their respective roles: Director (speaker), Matcher A (old addressee), and Matcher B (new addressee).

Apparatus

The director's gaze was tracked using an ISCAN ETL-400 remote eyetracker (sampling at a rate of 60 Hz). The eyetracker was placed on the table in front of the director, whose head movement was restricted using a chair with a headrest. The experiment stimuli was displayed using two monitors; the director's monitor was a 19" LCD with a 4:3 aspect ratio, and the matcher's monitor was a 17" LCD also with a 4:3 aspect ratio. The director wore headphones so that they could be informed of the target identity without the matcher overhearing.

Materials

The experimental stimuli consisted of 42 sets of pictures. Each set consisted of a target object, a competitor object, and four unrelated filler objects (one of which replaces the competitor during the test phase of the experiment). Two of the sets were used for practice trials that were not included in the analyses. Of the 40 experimental sets, 16 included a conventional target, 16 had an unconventional target, and 8 sets had conventional targets that did not have a competitor. Each of the stimulus images were 200x200 pixel bitmaps with the object pictures set on a black background. Descriptions of the pictures used as conventional targets are given in the Appendix.

Procedure

The experimenter began the experiment with a pre-recorded audiovisual presentation that introduced participants to the task. An example display containing five objects was shown to the participants, and the director was informed that his/her task was to describe one of these, the target, so that the matcher can identify it. The director was informed that he/she was not allowed to identify the target by mentioning the number of the space it appeared in or by otherwise indicating its location. The matchers' task was described as simply to select the target with the mouse. In the trials we called the "Feedback Available" trials, the matcher would be able to start moving the cursor immediately, and that participants would be able to freely converse. In the trials we called the "Feedback Unavailable" trials, matchers were told that they were not allowed to talk to the director, nor to move the mouse until the director had finished his/her description. For Feedback Unavailable trials, the director had to to indicate that he/she had finished the description by pressing a button on a response pad before any feedback would be made available. Triad participants were informed that there would only be one matcher present in the room for a given trial, and all participants were also reminded that "the

matcher who is outside of the room is not able to hear or see what is going on inside the room." They were also told to "take care to remember who is Matcher A and who is Matcher B." In the sessions involving triads, the matcher not participating in the current block waited out of earshot in an adjacent room until it was their turn to participate.

At the beginning of a trial, a display appeared on the screen showing the numbers 1 to 5 arranged in the locations where the pictures would eventually appear. After viewing this screen for 1000 ms, speakers heard a pre-recorded voice through headphones that announced the location of the target. 2000 ms after this, the numbers were replaced with five objects. In this way, speakers' gaze would tend to be on the target item when the objects appeared, and the number of shifts away from the target could be taken as a measure of the use of perceptually co-present information.

On Feedback Unavailable trials, speakers gave their description and pressed a button on a gamepad controller to indicate they had completed the description to their satisfaction. Prior to the button event, the addressee's mouse cursor was frozen in the middle of the screen. The addressee was only able to move the mouse after the button had been pressed. Addressees were not allowed to talk to the speaker on these trials. On Feedback Available trials, addressees could freely interact with the speaker. Also, the speaker did not need to press the button to indicate completion, and the listener's mouse cursor could be moved from the onset of the main display.

The experiment was divided into four blocks of trials, each of which in turn was divided into a training and a test phase. An example block of trials is given schematically in Figure 1. In the training phase of each block, speakers referred to two conventional and two unconventional targets presented five times each in a random order, thus establishing descriptions for these referents. In addition, they referred to a filler target during five additional filler trials during the training, for a total of 25 trials per training block.

Just prior to the test phase, there was a "transition phase" during which the

director was presented with a 30 second countdown until the start of the test phase, which gave the experimenter the opportunity to swap Matcher A for Matcher B and vice versa as necessary when there were multiple matchers. The countdown also served to give the directors in the non-partner switching condition an equivalent amount of downtime between the training and test phases. At the end of the countdown, the participants were presented with a screen advising them as to whether or not feedback would be allowed in the subsequent phase. Note that the feedback-availability manipulation was blocked for a given test phase (i.e., all trials within the phase were in the same feedback condition), and that speakers knew before going into the test phase whether or not the addressee was allowed to give feedback before speakers completed their descriptions.

After the transition events, the test phase began. In this phase of each block, speakers once again referred to the two old conventional and two old unconventional referents, in addition to two new conventional and two new unconventional referents. There were also two filler displays. The displays appeared in a random order. Unlike the training phase, when conventional referents appeared as targets, they appeared without the same-category competitor. The test phase displays appeared in a random order.

Design and Analysis

The analyses examined response measures during test trials, testing the main effects and interactions of Addressee (New vs. Old), the Novelty of the referent (New vs. Old), and Feedback Availability (Available vs. Unavailable). All of the analyses were performed using linear mixed effects models with by-subject and by-item random effects (Baayen, Davidson, & Bates, 2008). For each model, the link function was chosen to reflect the distribution of the dependent variable: poisson for count data (number of words, number of gaze shifts), binomial for binary data (whether or not a reference was overspecified), and gaussian for continuous data (log-transformed speech onset latency). Our models used

the "maximal" random effects justified by the experimental design (Barr, Levy, Scheepers, & Tily, under review): random intercepts for subjects and items, as well as random slopes for all treatment factors administered within a sampling unit (i.e., subject or item), and for all interactions involving factors that were administered within a sampling unit. For the current experiment, this meant by-subject random slopes for Novelty and Feedback Availability and their interaction, and by-item random slopes for Addressee and for Novelty. Models were fit using the `lmer` function of the `lme4` package version .999375-33 (Bates & Maechler, 2010) of the statistical software platform R. Models were estimated using maximum likelihood (ML) estimation rather than the default REML.

We used a model comparison approach to perform hypothesis tests. For each main effect or interaction of interest, we compared the deviance of a full model containing all fixed and random effects to that of a comparison model in which only the fixed effect being tested had been removed. All random slopes as well as higher-order fixed-effect interaction terms associated with that effect remained in the model. Thus, all test statistics followed the chi-squared distribution with one degree of freedom.

Although linear mixed-effects models allow simultaneous generalization to subjects and items, one drawback is that models capturing all of the dependencies in the random effects will sometimes fail to converge, generally attributable to difficulties in estimating the random effects. To goal of the modeling procedure was to come as close as possible to the maximal random-effects model, progressively removing from the model the random slopes, and/or the correlations associated with them, until the model converged. When choosing which random slopes/correlations to remove, we inspected the estimates from the partially converged model, and removed the highest-order slope that was associated with the least amount of variance (Barr et al., under review). If removing the high order slopes was not successful, we then considered correlations between the random slopes that were close to 1 or -1. We first followed this procedure with the "full model", and once we found

a model that converged, the resulting random effects structure was copied into all comparison models to ensure that the models differed only in the fixed effect of interest. If the comparison model required further removal of random slopes or correlations to achieve convergence, we fit a new full model that had an identical random effects structure to the comparison model.

Results

Overall, pairs were quite successful, with listeners clicking on the incorrect referent only 22 times out of 1024 total test trials (a 2.2% error rate). We removed four trials from the analysis that had an RT of zero (possibly due to an error in the experimental software). Below, we present results for unconventional and conventional referents separately, as they are intended to address different hypotheses (about underspecification and overspecification, respectively). Inferential statistics are presented in Table 1, and descriptive statistics are plotted in Figure 2.

[- - Table 1 goes about here - -]

[- - Figure 2 goes about here - -]

Unconventional referents

We first consider the effects of addressee, novelty, and feedback availability on description lengths for unconventional referents. For these referents, if speakers are egocentric and rely wholly on the abbreviated descriptions they have established for old referents, they risk producing underspecified descriptions that could create a misunderstanding or make interpretation difficult for the addressee. Our primary measure of underspecification, then, was the effect of Novelty: that is, the difference in description length for old versus new referents. A completely egocentric speaker would use as few words to describe an old referent for a new addressee as they would for an old one;

conversely, 'ideal speakers' who are maximally sensitive to their addressee's informational needs should describe an old referent to a new addressee using just as many words as they would have used had they been describing a new referent to that same addressee.

The length of descriptions (in words)[1] that speakers gave varied with their partner's knowledge (Addressee-by-Novelty interaction, $\chi^2(1) = 36.16$, $p < .001$). There was little evidence that speakers underspecified old referents: they used 55% longer descriptions when talking about these referents to new addressees compared to old addressees (8.3 versus 5.4 words; $\chi^2(1) = 9.21, p = .002$), replicating Wilkes-Gibbs and Clark (1992). There was still some evidence that speakers were influenced by their own knowledge, as the mean description lengths for new addressees and old referents (8.3 words) was, on average, about 11% shorter than that for new addressees and new referents (9.3 words), $\chi^2(1) = 5.31$, $p = .021$. One surprising result was that the Addressee-by-Novelty interaction was driven not only by a partner effect for old referents, but also by a partner effect for *new* referents: speakers used about 20% more words when speaking about new referents to old versus new addressees (11.5 vs. 9.3 words; simple effect of Addressee, $\chi^2(1) = 5.31$, $p = .021$).

The next question to consider is whether speakers managed to avoid underspecification through their own initiative, or because they were prompted to give longer descriptions by feedback from the new addressee. To address this question, we examine the effect of feedback availability. To the extent that speakers gave longer descriptions of old referents to a new addressee in reaction to overt signals of comprehension difficulty, then the effect should be stronger when feedback was available then when it was not. To the extent the adaptation was spontaneous, it should not depend on the availability of feedback; in other words, the relationship between addressee and novelty should be the same whether or not feedback was available. Supporting the idea that this adjustment was self-prompted, the three-way interaction was not reliable,

$\chi^2(1) = .01$, $p = .949$, with speakers lengthening their descriptions by about 61% (from 4.3 to 7.0 words) when feedback was available, and 51% when it was not (from 6.4 to 9.7 words). Although feedback availability did not seem to modulate speakers adjustments' to addressees, it is worth noting that speakers were sensitive overall to feedback availability, giving descriptions that were about 35% longer when feedback was unavailable (9.9 vs. 7.3 words; main effect of Feedback Availability: $\chi^2(1) = 6.51$, $p = .011$).

Clearly, the fact that speakers gave longer descriptions of old referents to new addressees when feedback was unavailable could only be explained as a self-prompted adaptation. However, this does not imply that speakers also adjusted to the addressee's needs spontaneously when feedback *was* available. It is also possible that speakers adapted their production strategies to the availability of feedback, allocating extra effort to planning and/or monitoring when feedback was unavailable, and relying more on collaborative processes when feedback was available. To test this possibility, we further analyzed the lengths of the descriptions speakers gave when feedback was allowed as a function of listeners' mouse feedback.[2] We measured the lag between the onset of the speaker's description and the first sustained movement of the mouse toward the target (with the criterion for a "sustained" movement being one that lasted at least 100 ms). If the lengthening of descriptions for old referents and new addressees is partially other-prompted, then the longer this "voice-to-mouse" lag time, the more evidence the speaker would have for the inadequacy of the current description, possibly motivating them to add more detail.

[- - Figure 3 goes about here - -]

We added the lag variable to the mixed-effects model, standardizing it so as to minimize effects of collinearity. We found a significant effect of lag, $\chi^2(1) = 10.05$, $p = .002$, such that for each standard deviation ($SD = 2050$ ms) above the mean lag time, the length of the description increased by approximately 1.4 words on average (Figure 3).

In fact, the voice-to-mouse lag time explained about 35% of the variance in the trial-level residuals from the full model.[3] This analysis strongly supports the view that speakers adapt how they gauge the informational adequacy of their descriptions to the potential for feedback, relying on feedback when available, and relying more on self-generated assessments of adequacy when feedback was unavailable.

We have seen that speakers can and do spontaneously avoid underspecification by lengthening their descriptions of old referents for new addressees. A further critical question concerns whether speakers achieve this through allocating extra effort to planning, or through monitoring and adjustment. To answer this question, we turn to the on-line processing measures of speech onset latency and visual scanning. To the extent that speakers chose to include more information during the initial design of their descriptions, this extra planning should have delayed the onset of speaking relative to speaking to the old partner.[4] There was no evidence for this; although the data showed the predicted interaction (Addressee-by-Novelty: $\chi^2(1) = 10.40$, $p = .001$), unexpectedly, the partner effect was reliable only when speakers described *new* referents (2465 vs. 1954 ms for old and new addressees, respectively; $\chi^2(1) = 4.08$, $p = .043$) but not when speakers described old ones (means = 1468 vs. 1346 ms for new vs. old addressees, respectively; $\chi^2(1) = 2.44$, $p = .119$; see Figure 2). This finding, though unexpected, is consistent with the earlier finding that speakers gave longer descriptions of new referents to old (vs. new) addressees. It is intriguing that the longer descriptions that speakers gave for new referents and old addresses seem to have been at least partly planned, whereas those given for old referents and new addressees seem to be wholly the result of monitoring and adjustment.

Further insight into the nature of these partner adaptations can potentially be found in the eye gaze data.[5] As noted in the Introduction, speakers should rely more on physically co-present information (and thus, should show a greater rate of gaze shifts) when planning and monitoring descriptions old referents for new addressees. This increase

in the use of visual context could be reflected either in planning (pre-onset gaze shifts) or in monitoring (post-onset gaze shifts). Overall, speakers did scan less when they planned descriptions of old vs. new referents (.9 versus 1.6 pre-onset gaze shifts respectively, main effect of Novelty: $\chi^2(1) = 7.56$, $p = .006$); however, there was no evidence that the scanning was partner-specific: although the Addressee-by-Novelty interaction was marginally reliable, $\chi^2(1) = 2.97$, $p = .085$, there was no evidence that speakers relied more on physically co-present information when planning descriptions of old referents for new addressees (means of .95 and .95 for old and new partners respectively, $\chi^2(1) = .06$, $p = .810$).

These considerations are partially qualified by a further (marginally significant) finding; namely, that speakers' sensitivity to their partners' needs during planning depended upon the availability of feedback (Addressee-by-Novelty-by-Feedback Availability interaction: $\chi^2(1) = 2.93$, $p = .087$). As indicated by Figure 2, speakers showed greater sensitivity to their partners' needs when feedback was unavailable (simple interaction of Addressee-by-Novelty: $\chi^2(1) = 7.43$, $p = .006$) than when feedback was available (simple interaction of Addressee-by-Novelty: $\chi^2(1) = .15$, $p = .702$). However, even in the feedback-unavailable case, in which stronger partner effects were present, there was still no evidence that speakers scanned more when describing old referents to new (vs. old) addressees (1.1 vs. .8 gaze shifts, respectively; $\chi^2(1) = 1.01$, $p = .315$). Instead, the interaction in the feedback-unavailable case seems to have been driven by speakers scanning the display more extensively when planning descriptions of *new* referents for the *old* addressee, although the effect was marginal (1.9 vs. 1.4 gaze shifts; $\chi^2(1) = 2.81$, $p = .094$).

There was more reliable evidence for audience design effects in the rate of post-onset scanning, which reflects monitoring and adjustment. For this analysis, we considered only the feedback-unavailable condition, up to the point at which the speaker pressed the

button. It is not sensible to analyze the data in the feedback-available condition, given the difficulty of distinguishing scanning related to considering display items versus scanning that reflects speakers following listeners' mouse movements. Speakers exhibited a stronger decline in scanning from new (vs. old) referents when they spoke to partners who shared their knowledge (about 50%; from 1.9 to .9) versus when they spoke to partners who did not (about 5%; from 1.4 to 1.3; Addressee-by-Novelty interaction: $\chi^2(1) = 7.59$, $p = .006$). But again, the key prediction that they would scan more when describing old referents to new partners was not supported ($\chi^2(1) = .18$, $p = .675$); nor was there evidence for a partner effect for new referents (1.9 vs. 1.4; $\chi^2(1) = .84$, $p = .361$).

Thus far, all the evidence we have reviewed offers no evidence that speakers avoided underspecifying old referents through additional planning, which implies that these adaptations were accomplished through monitoring and adjustment. If this hypothesis is correct, we might be able to find positive evidence for this in a content analysis of the descriptions. Specifically, monitoring and adjustment predicts that a substantial number of descriptions of old referents to new addresses should be of the form *"remembered expression (+ optional elaborated material)"*. We calculated the rate of such *elaborated repetitions* by comparing the description given for referents in the test phase to the terminal description given for that same referent in the training phase. This analysis revealed that elaborated repetitions were highest when speakers spoke to new addressees and feedback was available (Table 2). Between 28% (feedback unavailable) and 44% (feedback available) of all speakers' descriptions to new addressees were either repetitions or elaborated repetitions; this is compared to between 66% (feedback unavailable) and 73% (feedback available) for descriptions given to the same addressee.

In sum, when speakers talked about old referents, they successfully avoided underspecification, adapting to new addressees' needs for additional ~~for~~ information through monitoring and adjustment. When feedback was unavailable, they relied on

self-generated assessments of these needs. When feedback was available, they minimized these self-generated assessments by relying on addressee feedback. There was no evidence that producing longer descriptions of an old referent for a new addressee required additional planning. Unexpectedly, audience design effects on planning were only observed for speakers' descriptions of *new* rather than *old* referents: speakers gave longer descriptions of new referents for old (vs. new) addressees, taking about 400 ms longer to plan what they would say, and scanning the physical context more. These effects were not predicted, however, and so should be confirmed in later studies. We consider some possible interpretations in the General Discussion.

Finally, there was ample evidence that speakers took into account the interactional affordances of the situation, speaking longer and spending more time planning when feedback was unavailable. This is the first evidence we know of that interactional affordances of the conversational setting influence speech planning processes. However, there was little evidence that feedback interacted with audience design processes, apart from a marginally significant influence on pre-onset gaze shifts.

Conventional referents

Speakers managed to avoid underspecifying referents, but were they able to avoid overspecifying them? Before considering the results, it is important to note that speakers would only have had the opportunity to overspecify during the test phase if they had used specific terms during the training phase. Even though all training trials for conventional referents included a pair of objects from the same category—a target that was typical of the lexical category (e.g., an average-looking candle) as well as a competitor that was less typical (e.g., a melted candle)—on those trials speakers sometimes used only a bare noun to identify the target, e.g., calling the candle that was not melted "the candle". Because such cases created no opportunity for overspecification at test, we removed any test trial

for which the speaker had used the bare noun for the target on the training trial just prior to the test trial. This led to the elimination of 43 out of 480 trials (9%). We also eliminated test trials where speakers miscategorized the target (e.g, calling a red candle the "gas can" or the "red cylinder") or otherwise described the target in a way that would not apply to the competitor (e.g., calling the adult gorilla "King Kong"). This resulted in the removal of an additional 11 trials (2.3%). The final data set therefore contained a total of 427 out of 480 possible observations (89%). Results are given in Figure 2 and Table 1.

Overall, on the test trials, where speakers described old targets in the absence of any competitor, they overspecified the referent about 68.9% of the time; e.g., they called the typical candle "the unmelted candle". This was much higher than the 8.2% baseline level of overspecification for new targets (Novelty: $\chi^2(1) = 30.54$, $p < .001$). If speakers took the addressee's knowledge into account, then they should have been less likely to overspecify the referent when speaking to a new partner. However, overspecification rate seemed to depend only on the speaker's experience, but not the identity of the addressee. With the old partner, speakers overspecified old referents at a rate (76%) that was about 9.5 times higher than baseline (8%), whereas with the new partner, they did so at a rate (62%) that was about seven times higher than baseline (9%); however, there was little evidence that the increase relative to baseline differed across addressee (Addressee-by-Novelty interaction: $\chi^2(1) = .84$, $p = .360$). Whether or not speakers overspecified did not appear sensitive to the availability of feedback (ps $\geq .438$ for all main effects and interactions associated with this factor).

Consistent with the absence of robust audience design effects in the production measure, none of the processing measures showed any evidence for audience design. The time that speakers spent planning[6] was influenced only by the novelty of the referent, with speech starting 172 ms faster on average before an old than a new referent (Novelty: $\chi^2(1) = 6.73$, $p = .009$). There were no significant effects whatsoever related to pre-onset

or post-onset gaze shifts.[7] However, it is possible that the gaze shift measures were near floor, since there does not seem to be even a trend toward fewer gaze shifts in the old referent condition, as would be expected.

Unlike for the unconventional referents, there was little evidence that speakers managed to tailor their descriptions of conventional referents to addressees' informational needs. At first blush these results seem to conflict with those of Brennan and Clark (1996). However, they are actually quite consistent, because speakers in the Brennan and Clark (1996) study did not adapt immediately after the partner switch. When Brennan and Clark averaged over the four "trials" in their test block (wherein each trial involved reference to the all the objects in the test set), speakers showed greater reversion to unadorned noun phrases with new partners (going from "loafer" to "shoe"). However, an analysis that considered reversion on a trial-by-trial basis made it clear that the adaptation only emerged gradually, with no partner effect on the first of the four trials. This implies that speakers either learned the inadequacy of their responses (perhaps through feedback from the addressee, or by monitoring the adequacy of their own responses), or the memory traces from the training phase decayed sufficiently for speakers' long-term naming patterns to reassert themselves.[8] In our experiment, there were only references to two old conventional referents in each test block, and thus little opportunity for speakers to get the feedback from the listener that might cause them to adjust their descriptions. Given these considerations, our current results are fully consistent with those of Brennan and Clark. Moreover, our results show that speakers are prone to overspecify referents even when the overly specific terms are highly anomalous in a neutral context, such as calling a prototypical candle "the unmelted candle."

One question is whether speakers' overspecifications actually impaired comprehension, given previous suggestions that they can sometimes facilitate visual search (Arts et al., 2011; Mangold & Pobel, 1988). To test this, we examined listeners' response

latency[9] (time to click the target) as a function of whether or not speakers overspecified the referent. It only made sense to examine response times for descriptions of old referents in the feedback-available condition, since in the feedback-unavailable condition listeners were only able to move the mouse cursor after speakers pressed a button; such responses were therefore done "offline." We measured the latency to click the target from the onset of the noun (e.g., "candle" in "unmelted candle"). Listeners were indeed about 202 ms slower overall to click the target after hearing an overspecified description relative to a bare noun (4423 vs. 4221 ms, respectively; $\chi^2(1) = 5.79$, $p = .016$). Surprisingly, there was no evidence overspecified descriptions impaired comprehension more when addressees were unaware of the established description (161 ms vs. 243 ms penalty for new vs. old addressees; $\chi^2(1) = .09$, $p = .759$). It is not clear why this is the case; one possibility is that even though informed addressees had heard the description "unmelted candle" multiple times, they had always heard it with contextual support, and perhaps found the absence of such support anomalous.

Discussion

This article explored the interplay between attention and memory in language production, as a means toward developing an expert-performance model of audience design. The experiment provided new insight into these questions, by considering both over- and under-specification, by manipulating feedback to distinguish self-prompted from other-prompted adaptations, and by using online measures to gain insight into prearticulatory and postarticulatory processes of adaptation. To summarize our main findings:

1. Speakers overspecified referents at a very high rate, but rarely underspecified them (see also Deutsch & Pechmann, 1982; Ferreira et al., 2005);

2. As speakers accumulated experience describing abstract referents, they shortened

their descriptions (see also Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964), but were still able to adapt and lengthen them when describing these same referents to new (uninformed) addressees; Wilkes Gibbs

3. This lengthening was achieved incrementally (through self- and other-monitoring) and did not incur additional planning;

4. Speakers flexibly modulated how they assessed the informational adequacy of their utterances depending on the interactional affordances of the referring context, relying on feedback when available and self-generated assessments when unavailable;

5. Unexpectedly, speakers gave longer descriptions of *new* referents to *old* addressees, and spent more time planning these longer descriptions;

6. When speakers became accustomed to using modified descriptions for familiar referents (e.g., calling a candle "the unmelted candle" to distinguish it from another candle in the display), they persisted in using these modified descriptions even in contexts where the modification was no longer required (e.g., the contrasting candle was no longer present in the context), resulting in overspecified descriptions (see also Brennan & Clark, 1996);

7. However, speakers overspecified familiar referents at similar rates for new (uninformed) addressees as for addressees who shared the speaker's experience, supporting an explanation in terms of speaker-internal factors;

8. Addressees experienced greater difficulty comprehending overspecified references than adequately specified references (but see Arts et al., 2011; Mangold & Pobel, 1988).

Speakers overspecified referents at a very high rate, but rarely underspecified them (at least as measured by the number of words produced). Wilkes-Gibbs and Clark (1992) had shown previously that speakers gave longer descriptions of old referents to new addressees (and Galati & Brennan, 2010 showed that speakers gave longer descriptions of narrative events to new vs. old addressees). However, our findings go beyond these results by clarifying the role of feedback as well as by considering processing measures. Our

results indicate that the lengthening of descriptions for new addressees cannot be wholly explained by speakers adding more information in reaction to feedback from addressees. Without feedback, speakers spontaneously elaborated upon remembered expressions. The online measures also give further evidence that the adaptation is achieved through monitoring and adjustment, rather than being pre-planned.

A compelling explanation for the asymmetry between under- and over-specification is that speakers adapt their expressions incrementally (Pechmann, 1989). Speakers can avoid underspecifying referents because the cascaded nature of planning and articulation in language production allows them to incrementally add material to established descriptions in order to meet addressees' needs. Speakers can easily add more content to a reduced description (e.g., turning "the spiky shape" into "the spiky shape with the points coming out of the top") without incurring additional planning. This explains why speakers were able to lengthen their descriptions of old referents by about 55% for new addressees, but still managed to begin articulating their descriptions as quickly as they did with old addressees. In contrast, the incremental nature of production does not enable them to avoid overspecification, as one cannot incrementally delete or change material that has already been produced (except, one might argue, through the production of a repair, e.g., "the unmelted candle; uh, I mean the only candle"). In support of incrementality, none of the online measures showed any evidence for audience design effects during planning when speakers described old referents to new addressees.

One alternative explanation for the asymmetry between underspecification and overspecification is that perhaps the former is more communicatively costly than the latter. With too little information, listeners are forced to guess at a speaker's meaning; with too much information listeners might be able to identify the speaker's meaning but derive some additional (unintended) implicature. The risk of misunderstanding would seem to be higher in the former case. However, in interactive contexts, addressees can

always request more information until reference is successful, whereas it may be difficult and costly to "undo" an unintended implicature. Additionally, recent evidence suggests that increasing the cost of overspecification (e.g., incurring a financial penalty for "leaking" information that may be useful to a competitor) may actually ironically *increase* the likelihood of overspecification (Wardlow Lane & Liersch, in press). Thus, the evidence seems more consistent with an explanation in terms of incremental processing than in terms of communicative cost.

A further question is whether the failure to detect any evidence for audience design during the planning of descriptions for old referents might reflect a lack of power. This cannot be entirely the case, because audience design effects were in fact observed during planning, but only for unconventional referents, and only when speakers described *new* referents to *old* addressees. Speakers tended to use more words to describe new referents when speaking to the old versus the new partner, and also delayed more before they began speaking. There was also a marginally significant trend for speakers to scan the display more extensively when planning their utterances (but this was only significant when there was no feedback). What might this unexpected pattern reveal about audience design?

Given the unpredicted nature of these findings, we can only offer some tentative interpretations. One possibility is that speakers are wary about how they describe new referents to old addressees because they must produce descriptions that are lexically differentiated from previous descriptions, in order to avoid ambiguity (van der Wege, 2009). Consider a hypothetical example in which speaker A and addressee B have settled on calling referent X "the spiky shape," and then A is confronted with describing new referent Y. If Y resembles X, then A might be tempted to described it as being spiky, but to avoid confusing B, A would need to contrast that description with that already given for X (e.g., "a different spiky shape"). When talking to C, A would not need to do this. This could explain why descriptions of new referents were longer when delivered to old

rather than new addressees, why speakers took more time planning, and possibly why they might have done more scanning (i.e., to check the planned description against established descriptions for other referents in the array).

Although this explanation is appealing, it seems necessary to account for why speakers engaged audience design processes during planning for new referents but not for old ones. An intriguing possibility is that speakers use the strength of the memory signal as a cue to determine how much effort they allocate to planning. When speakers attend to a referent in order to describe it, previous descriptions are retrieved as an increasing function of the similarity of the current context to the previous contexts in which those descriptions had been used. This is essentially the main claim of the memory-based view of audience design (Horton & Gerrig, 2005). But in addition to this, we suggest that speakers exploit the strength of the signal (i.e., whether a single expression becomes strongly accessible, vs. multiple competing descriptions, vs. none at all) as a heuristic for allocating effort to utterance planning. When the signal is strong, that indicates to the speaker that the description is probably contextually adequate, causing the speaker to allocate little effort to planning. When the signal is weak, speakers allocate additional effort to planning and audience design. In this way, speakers can take advantage of encoding specificity to "automatically contextualize" their descriptions, and only incur additional planning effort when it is most likely to be needed. Although these ideas are consistent with our data, they are admittedly post hoc, and therefore warrant further investigation.

One important contribution of this work is that it is the first (to the best of our knowledge) to use feedback to clearly separate self-prompted from other-prompted adaptations. It is only the former category of adaptations that we feel should fall under the rubric of audience design. When speakers lengthened their descriptions of old referents for new partners in situations where feedback was disallowed, they did so spontaneously,

without requiring prompting from the partner. An additional novel finding is that speakers spent more time planning their descriptions of referents when they knew they could not rely on feedback, but only for unconventional referents. It is not clear why this did not hold for conventional referents; perhaps speakers view feedback as critical only in those cases where a conventional label for a referent is unavailable.

In general, our findings harmonize with the view that language users build up "routines" for producing and comprehending language in dialogue that short-circuit many of the complex layers of language processing (Pickering & Garrod, 2005). Effectively, these dialogue routines can be viewed as a kind of short-term expertise acquired through the effort to solve the various referential problems that recur over the course of a conversation. As repetitions accumulate, descriptions of objects stop behaving like descriptions and begin to behave more like proper names (Carroll, 1980); i.e., as rigidly denoting a particular object and operating in a manner that is progressively less sensitive to the referring context.

We feel that a view of audience design in terms of expert performance would be profitable for the field in that such a framework brings together issues of representation and processing. Representations are molded through experience in a domain, and attentional processes determine how these representations are adapted to solve problems that are similar, but not identical, to previously encountered problems. We have suggested the ITA model (Logan, 1988) as a particularly appealing framework. Our findings give us some new insights into how the ITA might be applied to audience design. First, the ITA assumes that the retrieval of stored solutions is an obligatory consequence of attention to the problem conditions. When speakers activate the intention to refer to a given referent, past solutions are retrieved obligatorily, depending on the similarity of the current retrieval context to previous contexts. We hypothesize that the strength of the memory signal is used to gauge the need for additional planning before starting to speak. If the

signal for a given solution is particularly strong (e.g., because the speaker has called this particular referent "the unmelted candle" five times in the recent past), articulation begins before audience design considerations are taken into account. This would explain why speakers in our experiment overspecified referents at nearly the same rate for new as for old addressees. It could also explain why audience design effects were found for speakers' descriptions of new referents: because no strong memory trace existed for these referents, speakers allocated substantial time for planning, and the need to lexically differentiate there descriptions from other descriptions to an old addressee required extra time.

Another aspect that should be taken into account in the development of such a model is the role of self- versus other-monitoring. Some of our results suggest that speakers can adapt how they monitor the informational adequacy of their descriptions to the interactional affordances of the referring context. When feedback is available, speakers tend to rely on cues provided by the partner; when such cues are unavailable, they rely on self-generated assessments. Furthermore, when speakers know that they cannot rely on feedback, they take more time planning their utterances. Models of language production currently do not account for this flexibility in how speakers plan and monitor their expressions.

References

Anderson, J. R. (1996). *The architecture of cognition.* Mahwah, N.J.: Lawrence Erlbaum.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*, 361–374.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, *28*, 937–962.

Barr, D. J., & Keysar, B. (2006). Perspective taking and the coordination of meaning in language use. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics (2nd ed.)* (pp. 901–938). Amsterdam, Netherlands: Elsevier.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (under review). Random effects structure in mixed-effects models: Keep it maximal. Manuscript under review.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using s4 classes [Computer software manual].

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, *79*, 941–952.

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, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology*, *19*, 441–472.

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

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*,

55–81.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*, 62–81.

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

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

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

Deutsch, W., & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, *11*, 159–184.

Engelhardt, P., Bailey, K., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, *54*, 554–573.

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

Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, *96*, 263–284.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*, 35–51.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination, and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*, 181–215.

Gerrig, R. J., & McKoon, G. (1998). The readiness is all: The functionality of memory-based text processing. *Discourse Processes*, *26*, 67–86.

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

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing *when* and *how* to adjust to addressees. *Journal of Memory and Language*, *47*, 589–606.

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

Keysar, B. (1997). Unconfounding common ground. *Discourse Processes*, *24*, 253–270.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*, 113–114.

Kraut, R. E., Lewis, S. H., & Swezey, L. W. (1982). Listener responsiveness and the coordination of conversation. *Journal of personality and social psychology*, *43*, 718–731.

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review*, *9*, 550–557.

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

Mangold, R., & Pobel, R. (1988). Informativeness and instrumentality in referential communication. *Journal of Language and Social Psychology*, *7*, 181–191.

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

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 89–110.

Pickering, M. J., & Garrod, S. (2005). Establishing and using routines during dialogue: Implications for psychology and linguistics. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones.* Mahwah, N.J.: Erlbaum.

Rawson, K. A. (2010). Defining and investigating automaticity in reading comprehension. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 52, pp. 185–230). Burlington: Academic Press.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*, 696–735.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*, 211–232.

Simon, H. A. (1996). *The sciences of the artificial.* Cambridge, MA: The MIT Press.

Steels, L. (1997). Self-organising vocabularies. In C. G. Langton & K. Shimohara (Eds.), *Artificial life v* (pp. 179–184). Cambridge, MA: MIT Press.

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

van der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*, 448–463.

Wardlow Lane, L., & Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1466–1481.

Wardlow Lane, L., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! *Psychological Science*, *17*, 273–277.

Wardlow Lane, L., & Liersch, M. J. (in press). Can you keep a secret? Increasing speakers' motivation to keep information confidential yields poorer outcomes. *Language and Cognitive Processes*, 1–12.

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*, 183–194.

Author Note

This work was presented at the 2011 Workshop on Language Production (Edinburgh). We thank the workshop participants who gave us feedback on our presentation. We also especially thank the undergraduate research assistants at UC Riverside who helped collect and transcribe the data: Amy Dell, Justin Do, Taiki Kondo, Susanna Luu, Adrian Molina, Ashley Rodolf, and Jessica Zendejas. Thanks also to Mandana Seyfeddinipur who helped supervise data collection and coding.

Footnotes

[1]To minimize the influence of outliers on the description lengths, we truncated observations at 26 words, the 97.5 percentile of the distribution.

[2]We thank an anonymous reviewer for the suggestion of analyzing the description lengths as a function of the actual feedback that speakers received. We focused on mouse feedback because we witnessed very little verbal exchange between speakers and listeners in this experiment, perhaps because the task was so easy.

[3]To calculate the trial-level residuals, we generated predictions from the mixed-effects model excluding lag (including estimated random effects offsets for subjects and items), and subtracted these from the observed values.

[4]Speech onset latency values were truncated at 4216 ms, the 97.5th percentile of the distribution.

[5]The number of pre- and post-onset gaze shifts was truncated at the 97.5th percentile of the corresponding distribution (5 and 9, respectively).

[6]We trimmed the speech onset latency distribution at the 97.5th percentile (2645 ms)

[7]The gaze shift measures were truncated at the 97.5th percentile of their distributions, which was 4 for both measures.

[8]Brennan and Clark (1996) do note that in the first trial of the critical test phase, speakers were more likely to abandon their conceptual pacts with a new partner. However, all that is really relevant here is whether the descriptions they used were overly specific or not. It is possible for a speaker to have switched from one subordinate term to another (e.g., from *sneaker* to *tennis shoe*), rather than to have switched from an overly specific term to a basic-level term (e.g., from *sneaker* to *shoe*). Although both cases reflect an abandoning of the conceptual pact, it is only the latter of these that could truly be characterized as an accommodation to the addressee's less-informed perspective.

[9]Values were truncated at 6473 ms, the 97.5th percentile of the distribution.

Appendix: Conventional Items

Target	Competitor	Modal Response
candle	(not melted) candle	unmelted candle
key	(old-fashioned) key	modern/new/gold key
knife	(Swiss Army) knife	knife with the brown/wooden handle
trash can	(metal) trash can	plastic/white trash can
spoon	(large slotted) spoon	small spoon
guitar	(electric) guitar	acoustic guitar
carrot	(cartoon) carrot	little/real carrot
gorilla	(young/brown) gorilla	black gorilla
gun	(toy) gun	real gun
leaf	(dark green three pointed) leaf	dark green leaf
rose	(wilted; stem not visible) rose	rose with a stem
marker	marker (with cap off)	marker with a cap
screwdriver	screwdriver (with black on handle)	screwdriver, all red handle
backpack	(blue) backpack	purple backpack
clamp	(open) clamp	closed clamp

Table 1. Inferential Statistics.

	Unconventional Referents							
	Production Measures				Gaze shifts			
	Word Count		Speech Onset		Pre-Onset		Post-Onset	
Factor	$\chi^2(1)$	p	$\chi^2(1)$	p	$\chi^2(1)$	p	$\chi^2(1)$	p
Addressee (A)	.91	.340	1.04	.307	.86	.354	.45	.501
Novelty (N)	34.30	<.001	34.60	<.001	7.56	.006	12.50	<.001
Feedback (F)	6.51	.011	4.12	.042	.76	.383		
A:N	36.16	<.001	10.40	.001	2.97	.085	7.59	.006
A:F	.21	.649	.87	.351	3.79	.052		
N:F	.29	.589	.12	.726	.92	.336		
A:N:F	.01	.949	1.28	.257	2.93	.087		
	Conventional Referents							
	Production Measures				Gaze shifts			
	Oversp. Rate		Speech Onset		Pre-Onset		Post-Onset	
Factor	$\chi^2(1)$	p	$\chi^2(1)$	p	$\chi^2(1)$	p	$\chi^2(1)$	p
Addressee (A)	.27	.604	1.77	.183	2.49	.115	.13	.721
Novelty (N)	29.88	<.001	6.73	.009	1.21	.271	1.29	.257
Feedback (F)	.25	.618	.15	.694	.02	.891		
A:N	.84	.360	1.69	.194	.33	.567	1.13	.287
A:F	.01	.937	.27	.603	.26	.613		
A:N:F	.16	.693	1.90	.168	.01	.951		

Table 2. Analysis of Abstract Descriptions.

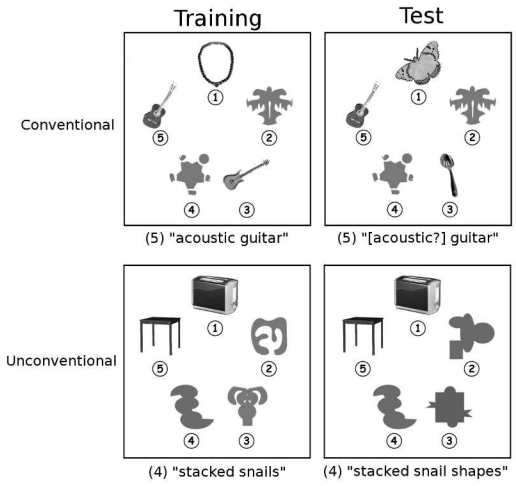
Addressee	Feedback Allowed	Repetition/ Reduction	Elaborated Repetition	Reconcept.
Old	Yes	73%	0%	27%
Old	No	55%	11%	34%
New	Yes	11%	33%	56%
New	No	13%	15%	72%

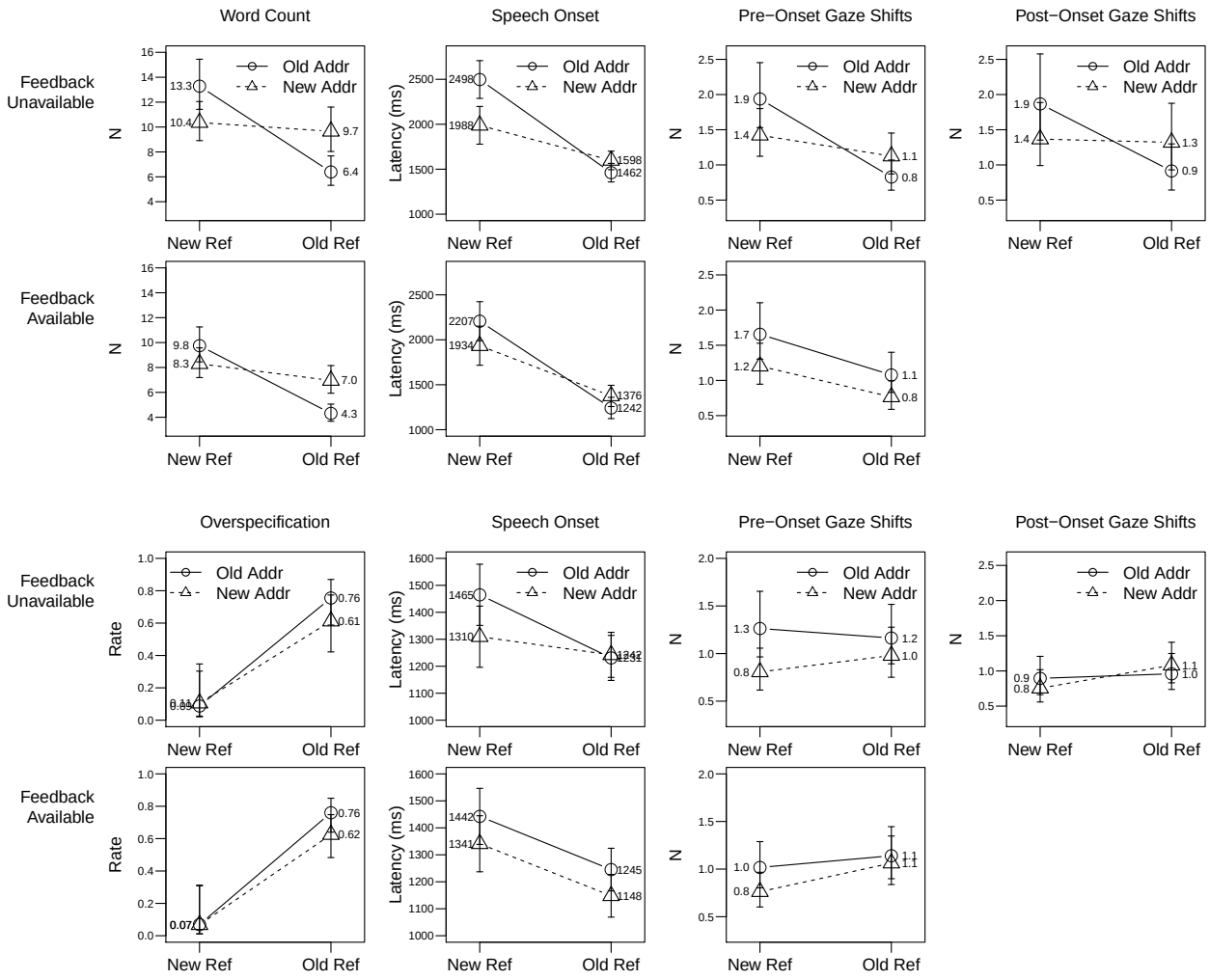
Figure Captions

Figure 1. Example training and test displays for conventional and unconventional referents. The text below each panel indicates the location of the target (by number) and provides a sample utterance.

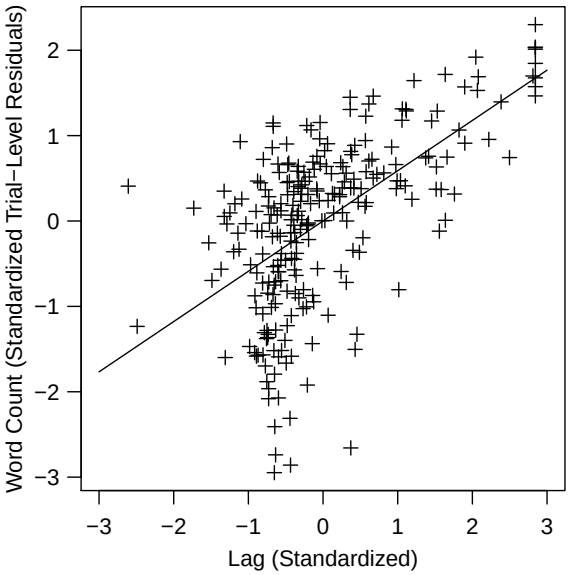
Figure 2. Descriptive statistics for conventional (top two rows) and unconventional (bottom two rows) referents.

Figure 3. Voice-to-Mouse Lag (horizontal axis) plotted against trial-level residuals of Description Length (vertical axis) from a model excluding lag.


Training
Test
Conventional
Unconventional
(5) "acoustic guitar"
(5) "[acoustic?] guitar"
(4) "stacked snails"
(4) "stacked snail shapes"


Word Count
Speech Onset
Pre−Onset Gaze Shifts
Post−Onset Gaze Shifts
Feedback Unavailable
Feedback Available
Old Addr
New Addr
N
Latency (ms)
13.3
10.4
9.7
6.4
2498
1988
1598
1462
1.9
1.4
1.1
0.8
1.9
1.4
1.3
0.9
9.8
8.3
7.0
4.3
2207
1934
1376
1242
1.7
1.2
1.1
0.8
New Ref
Old Ref

Overspecification
Speech Onset
Pre−Onset Gaze Shifts
Post−Onset Gaze Shifts
Feedback Unavailable
Feedback Available
Old Addr
New Addr
Rate
Latency (ms)
N
0.76
0.61
0.07
0.76
0.62
0.07
1465
1310
1442
1341
1245
1148
1.3
0.8
1.2
1.0
1.0
0.8
1.1
1.1
0.9
0.8
1.1
1.0
New Ref
Old Ref


Word Count (Standardized Trial−Level Residuals)
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Lag (Standardized)