Philosophy Compass 2/3 (2007): 475–496, 10.1111/j.1747-9991.2007.00082.x

Implicature During Real Time Conversation: A View from Language Processing Research

Julie C. Sedivy*
Brown University

Abstract

Grice's notion of conversational implicature requires that speaker meaning be calculable on the basis of sentence meaning, and presumptions about the speaker's adherence to cooperative principles of conversation and the ability of the hearer to work out the speaker's meaning. However, the actual real-time consideration of cooperative principles by both the hearer and speaker runs up against severe temporal constraints during language processing. This article considers the role of language processing research in the shaping of a theory of implicature, and provides an empirical overview of pertinent current work in real-time language production and comprehension.

1. Considering Theories of Implicature in a Cognitive Processing Context

A good deal of what is communicated takes place 'between the lines' of conventional meanings, relying on the speaker's and hearer's coordination of conversational expectations, and the juxtapositioning of these expectations with the conventional meaning of the utterance. For example, consider the following scrap of discourse:

1) Richard asked Elizabeth why she refused to marry him. She replied that a handful of her long string of marriages had ended with feelings of mutual respect.

Here, the speaker probably manages to convey a good bit of unspoken meaning, including the fact that only a small proportion of Elizabeth's marriages ended amicably, that she has low expectations for the chance of success of a marriage to Richard, and that the likelihood of an acrimonious separation is a valid reason to reject a proposal.

Grice's work on implicature has provided a useful framework for thinking about this important contribution to meaning by emphasizing the distinction between conventional and understood meanings, and sketching out a set of communicative principles through which understood meanings might be derived on the basis of conventional meanings. A critical feature of Grice's conception of conversational implicatures is the notion that they are calculable. Thus, a speaker who says *p* may implicate *q*:

> PROVIDED THAT (1) he is to be presumed to be observing the conversational maxims, or at least the cooperative principle; (2) the supposition that he is aware that, or thinks that, *q* is required in order to make his saying *p* . . . consistent with this presumption; (3) the speaker thinks (and would expect the hearer to think that the speaker thinks) that it is within the competence of the hearer to work out, or grasp intuitively, that (2) is required. (49–50)

Grice never explicitly connected his ideas to theories of real-time language production and processing, intending to provide an explanatory account of the mechanisms whereby implications *could* be conveyed, rather than a predictive account of what implications actually *are* conveyed, and under which circumstances. In order to specify how these mechanisms are actually used in real settings spanning a range of communicative situations, the emphasis on calculability requires that a theory of implicature be embedded in models of real-time language processing. For example, if the speaker's meaning is dependent on the assumption that the hearer can 'work out' the implication on the basis of the conventional meaning, then, prior to executing what is said (i.e. planning the sequence of words to be uttered and specifying their articulatory instructions), the speaker must somehow be able to compute what the hearer is likely to be able to work out on the basis of the conventional meaning. Conversely, in order for the implication to successfully go through, the hearer must actually be able to work out the intended meaning under the real-time pressures of that particular communicative setting. In conversation, this means being able to compute the implicature triggered by an expression or utterance in sufficient time that it does not interfere with the processing of subsequent incoming material that is part of a continuous, rapid stream of linguistic input.

Leisurely introspection of carefully crafted example sentences in scholarly articles masks the intensely frenetic nature of everyday real-time language processing. Over the last few decades, a large body of experimental work in language production and comprehension has thrown into relief the temporal demands on the language processing system. Conversational partners must cope with sifting through a huge quantity of interacting information in order to articulate or interpret an utterance. It is now known that both language comprehension and production are highly *incremental* in nature; that is, commitments to planning or interpretation are made on the fly on a moment-by-moment basis, relying on partial computation, and in the absence of full knowledge about the linguistic expressions that are yet to be uttered. In production, this means that a word is often being uttered before the processing system has completed the selection of words more than a word or two downstream, or before it has committed to a syntactic structure. In comprehension, hypotheses about the structure and meaning of an utterance are being initiated long before the hearer encounters decisive evidence that would exclude numerous incorrect hypotheses. Because of the systematic time gap between initiating the processing of an expression and the availability of all information pertaining to its processing, a great deal of the debate in the processing literature has focused on specifying the time course in which various kinds of information are actually used. Hence, one cannot take it for granted that all information that would in principle be pertinent for the design or interpretation of an utterance is available at just the point in time where the decision must be made. With regards to implicature, one needs to ask what information is available to enter in the calculations that the implicature depends upon.

Grice's goal of characterizing the aspect of meaning that he identified as implicature is clearly distinct from the aims of cognitively-oriented researchers who seek to give a characterization of the real-time psychological implementation of this aspect of meaning. However, theoreticians disagree amongst themselves about how sharp the separation between the two enterprises really should be. Some, such as Bach, Saul, and Horn argue that neither cognitive considerations nor considerations of how hearers interpret utterances should bear on a theory of implicature, which is properly thought of as part of an account of speaker meaning concerned with 'how and why the speaker, given what she wants to convey, utters what she utters' (Horn, 'Implicature' 194). Others, such as proponents of Relevance Theory (Carston; Sperber and Wilson), claim that the phenomena that are targeted by Grice's account of implicature really belong to a larger domain of inquiry which seeks to explain human communicative behavior within a cognitive perspective. Indeed, the main argument of Relevance Theory is that implicatures arise because speakers and hearers jointly assume that speakers will choose an utterance that strikes an optimal balance in providing hearers with maximal information for minimal cognitive effort. For relevance theorists, this mutually held assumption subsumes all of Grice's distinct maxims of quality, quantity, relevance, and manner, replacing Grice's rationalist account with an information-processing account. This is a radical shift, with the effect of moving an explanation of implicature away from the *personal* level of explanation, which deals with rational, norm-based agency, onto the *sub-personal* level, which deals with mechanisms (such as unconscious psychological processes) that play a causal role in behavior (Carston). Such a shift in orientation has far-reaching implications. As Carston points out,'set within a cognitive-scientific framework, this kind of pragmatic theorising is answerable to quite different sources of evidence and criteria of adequacy from any philosophical analytical investigation' (129). Indeed, in order to have any real predictive force, Relevance Theory critically depends on a detailed understanding of the cognitive processing costs of various inferences as well as an account of how speakers integrate awareness of these costs into their choice of linguistic expressions.

However, a mechanistic view of implicature is not limited to proponents of Relevance Theory. Levinson articulates a neo-Gricean perspective in which he mounts a vehement attack on Relevance Theory's attempt to reduce all of Grice's maxims to a single information processing principle. Nevertheless, he shares Relevance Theory's cognitive orientation, particularly in his justification of a distinction between what Grice termed *generalized* and *particularized* conversational implicatures. The distinction can be illustrated by reconsidering example (1), repeated below, in contrast with (2):

(1) Richard asked Elizabeth why she refused to marry him. She replied that a handful of her long string of marriages had ended with feelings of mutual respect.

(2) Richard asked Elizabeth why she'd agreed to marry him. She replied that a handful of her long string of marriages had ended with feelings of mutual respect.

The identical second sentences in (1) and (2) give rise to different implicatures, shaped by the preceding context sentence. In (1), the speaker is understood as implicating that the bitterness of most of her marital endings is seen by Elizabeth as a valid reason to decline marriage to Richard, whereas in (2), the same sentence conveys that the amicable endings of a subset of her marriages warrants giving marriage to Richard a shot (while communicating that an amicable divorce is likely the best that one could hope for from this marriage). The context-sensitive contribution of meaning that we see in the second sentence of (1) and (2) is said to reflect a *particularized* implicature (on the basis of Grice's maxim of relevance). However, *both* (1) and (2) carry the implicature that it is not the case that many or all of Elizabeth's marriages ended amicably. This has been argued to be due to a *generalized* implicature that is stable across contexts (though of course it may be cancelled), and arises out of the general expectation that the use of a weak expression (*a handful*) signals the negation of stronger alternative expressions (*many*, *most*, *all*, etc.). What is interesting about these cases is the possibility of computing a large class of implicatures fairly directly from their conventional meanings – all that is needed is a pre-existing scalar relationship of the target expression to accessible stronger alternatives, plus the activation of a conversational principle that presumes speakers are maximally informative wherever possible (adhering to Grice's quantity maxim). Information about the specific context and its relationship to the utterance need not be considered.

From a cognitive perspective, if a class of implicatures can be shown to have the properties of computational generality and robustness across contexts, such implicatures might place only relatively slight demands on the processing system, thereby allowing for systematic pragmatic enrichment of conventional meanings in a large range of conversational settings. According to Levinson, a strong argument for the existence of generalized conversational implicatures is that they provide a powerful means of getting around the bottleneck on communication speed that is imposed by the physiological constraints on the human articulatory system. His communicative heuristics (derived from Grice's maxims of quantity and manner)

> serve to multiply the informational content of any message by a factor of perhaps a score, transforming the slow coding rate of human speech into something

> approximating the speed of human communication. All that is required for such a system to work is a tacit agreement between communicators that such heuristics can be assumed to be operative unless there are indications otherwise. (34)

Levinson further argues for generalized implicatures partly on the basis of their compatibility with the incremental nature of the human processing system. For example, scalar implicatures of the sort described above can arise due to a set of lexically coded oppositions between expressions that are shared by the speaker and hearer in conjunction with general presumptions about what a speaker intends by the choice of one over another expression. This means that the implicature (e.g.'*a handful*' means '*not many*') can be computed by the hearer as soon as the pertinent expression is encountered, rather than waiting for the entire sentence to unfold such that something like a full proposition can be represented. Levinson, then, maintains Grice's distinction between generalized and conversationalized implicatures as well as many of the details of Grice's original maxims, while embedding these ideas within a resoundingly mechanistic view.

To the extent that one relies on cognitive facts to motivate pragmatic theory, a detailed understanding of the processing dynamics of implicature has the potential to bear on distinctions among competing theories of implicature. For example, the division between conversational and particularized implicatures figures prominently in several theories of implicature which place importance on the claim that, unlike particularized implicature, generalized implicatures are context-general and intimately connected with grammatical phenomena (e.g. Horn, 'Toward a New Taxonomy'; Levinson). Indeed, some researchers go so far as to claim that scalar implicatures are properly thought of as aspects of sentence meaning rather than speaker meaning (Chierchia; Davis). On the other hand, the generalized/particularized distinction is broken down by some theorists who argue that conversational maxims such as quantity, which typically yield generalized implicatures, interact and are interdependent with maxims such as relevance, which are heavily context-dependent (e.g. Matsumoto; Potts). Still others, such as the relevance theorists, have argued that so-called generalized and particularized conversational implicatures are both generated by one and the same context-dependent process (Carston,'Informativeness'; Sperber and Wilson).

This distinction is related to broader distinctions within pragmatic theory that have been seen as theoretically important. For instance, both Bach and Recanati draw a distinction between primary pragmatic processes that rely on broad contextual information but do not involve inference, and secondary pragmatic processes such as Gricean implicature which crucially rely on the identification of speaker meaning. Indeed, for Recanati, the distinction has cognitive consequences: Because primary processes are non-inferential and do not involve the computation of the literal meaning of the global utterance, local processing of enriched meanings is possible prior to the identification of the truth conditions of an utterance. On the other hand, implicatures, which are inferential, presumably do depend on the global computation of the literal meaning of the utterance. This view is quite different from Levinson's, in which he argues for an intermediate level of meaning between speaker meaning and sentence meaning. This level includes many of the 'enriched meaning' phenomena that Bach and Recanati take to be non-inferential, as well as generalized conversation implicatures, all of which are argued to be computed locally and incrementally. There is, therefore, considerable and fundamental disagreement among researchers about the appropriate way to classify certain pragmatic phenomena initially identified by Grice as implicature. If different types of pragmatic processes could be identified as having distinct processing 'signatures', such work could potentially bear on these theoretical disagreements.

While more classically oriented neo-Griceans maintain a greater distance from empirical psychological evidence, it is not hard to see how certain cognitive facts could in principle either challenge or bolster a Gricean account of implicature. For example, a central goal of Grice's program is to argue that one can sidestep potential semantic ambiguity of a range of expressions (including quantifiers, modal operators, logical connectives 'and' and 'or', etc.) by appealing to an enrichment of a single conventional meaning by means of implicature. Suppose, however, that it can be shown that recovering a speaker's meaning via the maxims is outside of the hearer's processing limitations under many normal conversational circumstances, and yet the hearer nevertheless manages to arrive at the 'enriched' meanings of certain expressions. This would compromise Grice's program, in the following way: While it might be possible for the *speaker* to arrive at his intended meaning without invoking ambiguity of meaning, the *hearer* would need to have access to multiple conventional meanings in order to arrive at the intended meaning under typical temporal pressures.

Furthermore, severe cognitive limits on the speaker or hearer's ability to integrate the conversational maxims in real-time have the potential to undermine the shared presumption that the speaker is behaving in accordance with the cooperative principle. A rationalist-based account of implicature loses some force if speakers are observed to abide by the cooperative principle only sometimes, but not reliably, and if hearers are able to work out the speaker's intended meanings on the basis of this principle only under some circumstances. If these facts were true, a speaker's adherence to the cooperative principle could not reasonably be taken as the default mutual assumption.

On the other hand, certain facts about cognitive limitations could conceivably alleviate some important problems associated with a Gricean account. For example, Davis points out numerous cases in which Gricean theory overgenerates implicatures, predicting certain implicatures that do not in fact arise. If implicatures are prevented from occurring in these circumstances due to cognitive or processing factors, the generality of the pragmatic account can be maintained, while explaining away the apparent counterexamples by appealing to facts about the cognitive implementation of the pragmatic principles.

In the remainder of this article, I will explore some aspects of the relationship between real-time language processing mechanisms and conversational implicature. It is important to note that systematic study of this relationship is still in its very early stages and that we are very far away from having a good general understanding of the nature of pragmatic processing. Furthermore, the studies within the language processing literature are typically more focused on addressing questions about the architecture and mechanisms of the processing system itself than on questions that might inform pragmatic theories. It would be premature at this point to claim that the field has decisive evidence that bears directly on the formulation of a theory of implicature. Nevertheless, it is possible to identify a number of questions and approaches from the processing literature that are of potential interest to theoreticians. In the remainder of this paper, I aim to briefly summarize some of this work.

2. Cognitive Processing Limits for the Speaker

Under Grice's account of implicature, speakers are presumed to choose their utterances in accordance with the four conversational maxims. However, the demands on the language production system may impose limits on the extent to which speakers may actually be able to adhere to the maxims. The bulk of the experimental work addressing this issue can be related to the maxims of manner and quantity.

Under the maxim of manner, the speaker is assumed to take ease of comprehension into account in utterance planning, so as to avoid unclarity, wordiness (which presumably taxes comprehension resources), and ambiguity.

Avoidance of ambiguity is a particularly interesting point. The incremental nature of language comprehension leads to a potential proliferation of ambiguity for the hearer, who may need to make commitments about structure and meaning at a point in time where the linguistic input is compatible with numerous alternatives, and prior to receiving upcoming disambiguating information. The following examples illustrate the point, showing two possible continuations of an ambiguous string that each rest on a very different structural analysis. Such ambiguities are well-documented to cause processing difficulty for the hearer (or reader) particularly for the second of each of the continuations.[1]

(3) The student examined . . .
(4) . . . the test tube before making notes.
(5) . . . by the committee was awarded her doctorate.
(6) The coach knew you . . .
(7) . . . since you were ten.
(8) . . . missed practice.

In many cases of such temporary ambiguity, there exists an alternative, unambiguous way to express one of the meanings associated with the ambiguous string. For example, if one wished to convey (5), one could unambiguously say '*The student who was examined* by the committee was awarded her doctorate', thereby avoiding the processing difficulty. Similarly, instead of (8), one could say '*The coach knew that you* missed practice'. These alternatives suggest that in principle, if a speaker were to choose the temporarily ambiguous versions such as (3) or (6), the hearer might in all fairness presume that the speaker must be intending to communicate the structure consistent with (4) or (7), because had the speaker intended the structure required for (5) or (8), he would have used the unambiguous form available. Such reasoning represents a provocative extension of Gricean theory into the problem of real-time ambiguity resolution, a central problem for language processing researchers. This account of course relies on the speaker's being aware of the difficulties for the hearer posed by the ambiguity, and the hearer's ability to recognize that an unambiguous alternative exists for one of the interpretations, along with the potential for calculability of the intended meaning on the basis of the Cooperative Principle.

Current experimental evidence suggests that such computations are typically beyond the speaker's limits during real-time utterance planning, at least with regards to considering the difficulty of a *particular* utterance for the hearer. Ferreira and Dell investigated how likely speakers were to include the disambiguating function word 'that' in completing ambiguous fragments such as (6) above, as compared to completing unambiguous fragments such as (9), in which case-marking precludes the possibility of a temporary ambiguity.

(9) The coach knew I . . .

They found that the potential for ambiguity did not affect the inclusion of *that*; rather, this depended on production-centered considerations, such as the salience or accessibility of the words being encoded. This occurred despite the fact the hearers expressed a clear preference for the disambiguated structures. Similar results were found by Arnold et al. in investigating speakers' choice of ambiguous sentences versus unambiguous syntactic orderings with the same meaning. Again, factors internal to the production system were found to determine the speaker's choice, regardless of the potential for disruption of comprehension for the hearer, and again, despite hearers' preference for the disambiguated structures. Indeed, grammatical means of disambiguation are frequently not even exploited in written language, where the time pressures on production are far less dire (Elsness).

Interestingly, some of Ferreira and Dell's findings suggest that speakers do show an adjustment of their utterances to a hearer, but can typically only do so in a coarse manner that generalizes across utterances, rather than by predicting the difficulty of a particular utterance for the hearer. When speakers addressed their utterances directly to a co-present hearer (rather than producing tape-recorded utterances for a non-present hearer), and the need for clarity was emphasized, speakers were more likely to use disambiguating grammatical function words; however, they did so across the board, regardless of whether the sentence contained an ambiguity in the absence of the function word.

Prosodic intonation, as well as the use of function words, has the potential to disambiguate. Because intonational pauses correlate with syntactic phrase boundaries, a syntactically ambiguous sentence can often be prosodically disambiguated, as illustrated below: (Here, '. . .' denotes a pause in the speech)

(10) Put the apple . . . on the towel in the box. (indicates the structure: 'Put [the apple] [on the towel in the box]' i.e. Put the apple on the towel that's in the box.)
(11) Put the apple on the towel . . . in the box. (structure: 'Put [the apple on the towel] [in the box]' i.e. Put the apple that's on the towel in the box.)

Is there evidence that speakers systematically use prosody as a way to reduce ambiguity for the hearer? As with the example of function words, this feat would rely on the speaker's capacity to detect potential problems for the hearer in sufficient time to affect the planning of the prosodic boundaries. Overall, the experimental evidence suggests similar limitations on speakers' ability to use prosody as have been observed for the production of function words: prosodic information does serve as a helpful cue to the hearer in avoiding ambiguity, but it often appears to be done without such specific intent by the speaker. For example, a detailed study by Kraljic and Brennan found that speakers marked prosodic boundaries regardless of whether this was needed to disambiguate a sentence, regardless of whether they themselves had experienced such potential ambiguity by first performing the addressee's task in the experiment, and regardless of whether they had simultaneously disambiguated the instruction through the use of a function word. Therefore, it appears that a speaker's decision to mark prosodic boundaries is also driven mostly by production-based constraints. Interestingly, however, the degree of prosodic marking does appear to be sensitive to the communicative consequences of speech, though in a highly general way. That is, while Albritton, McKoon, and Ratcliff found that untrained readers rarely used prosody to disambiguate structure, a highly interactive game-playing experimental task by Schafer et al. found that speakers liberally used prosody to disambiguate structure; however, they did so regardless of whether disambiguation was already possible through other means such as visually available context. It seems, then, that speakers implement some awareness of the potential usefulness to the hearer of prosody and function words in their speech in fluent conversation; however, they are typically so preoccupied with the demands on the production system that they make use of these in a coarse, and often redundant way, adopting a general strategy of increasing linguistic information without calculating the precise information that is most useful to the hearer at the time. This may be quite analogous to the impulse to speak loudly and slowly when addressing a child or foreign person. It should be noted, however, that one study (Snedeker and Trueswell) did find a relationship between speakers' use of disambiguating prosody and potential for ambiguity. In this study, the instructions generated by the speakers were simpler than in either the Kraljic and Brennan experiments or the Schafer et al. study. Hence, nuanced prosodic planning in anticipation of ambiguity may not be categorically precluded, but rather, is something that must frequently be sacrificed in the face of production pressures.

Further work looking at speakers' avoidance of ambiguity of words is consistent with the findings pertaining to ambiguity of structure. Ferreira, Slevc, and Rogers engaged speakers in a referential communication task with instructions to describe a set of pictures from a visual display in such a way that each picture could be uniquely identified by a potential hearer. The visual display for a number of the trials introduced a lexical (word) ambiguity. For example, the set of pictures might include a bat (flying mammal), as well as a foil object associated with a homophone (baseball bat). Are speakers able to anticipate and avoid the potential ambiguity arising from the homophones by providing additional modification of the ambiguous noun (e.g. 'the flying bat' or 'the bat with wings')? The study showed that they did so to a limited degree. They did add modification somewhat more often to nouns in situations where the display contained a homophone foil; however, this increase was quite modest. Interestingly, the physical presence of the hearer again resulted in global adjustments to the specificity of descriptions with speakers adding modification across the board, regardless of whether this would result in disambiguating reference, or adding redundancy to the description.

Speakers therefore seem to show some clear limitations to their capacity to anticipate and avoid the difficulty hearers might experience as a result of word or syntactic ambiguity. Hence, they manage to avoid syntactic or lexical ambiguity with only limited reliability. However, when it comes to the actual planning of the content of the utterance, it turns out that speakers show a striking ability to anticipate difficulty for the hearer, and to respond to the particulars of the context in doing so. In the same study, the homophone displays were compared with displays containing a referential foil (a picture of the same category as the target picture). For example, the speaker might see a display containing two flying bats, one of which was larger than the other, allowing for unique identification through the use of a modifier (e.g. 'the small bat'). For these displays, speakers showed exquisite sensitivity to potential referential indeterminacy[2] for the addressee, and essentially never produced a bare noun in displays containing a referential foil. This finding is consistent with many studies which have found that speakers very robustly provide sufficient information to uniquely identify a referent from among a set of potential referents, when the indeterminacy does not hinge on an ambiguity of form (see for example Deutsch and Pechmann; Engelhardt, Bailey, and Ferreira; Olson; Sedivy). Indeed, there is direct evidence for the remarkable speed with which speakers can perceive the potential referential indeterminacy for the hearer, and recruit production mechanisms in response. Brown-Schmidt and Tanenhaus tracked speakers' eye movements in a similar referential communication task, and found a tight time-locking between when the speaker looked at the referential foil, and the occurrence of a disambiguating modifier. If the foil was noticed on average 8/10 of a second or more before the beginning of the utterance, the speaker typically produced a fluent description involving a prenominal modifier. This is approximately the amount of time it generally takes to initiate an utterance upon viewing a picture of the referent. For example, it normally takes slightly less than a second between seeing a picture (say, of a couch), and beginning to utter 'couch'. The Brown-Schmidt and Tanenhaus results are quite striking, because they show that there is no appreciable difference in the time course of retrieving a simple noun label for a referent and determining the need for modification as a result of the presence of a referential foil. In avoiding referential indeterminacy, then, potential difficulties for successful reference using a bare noun are perceived very rapidly, reliably, and with consideration of the particulars of the context.

Ferreira et al. explain the contrast in their results between the disambiguation of form versus content by pointing to the different processing stages implicated in production (for a general review of production mechanisms, see Levelt). The decision to include a modifier to distinguish one referent from another takes place at a message formulation stage, at which point the semantic content of the utterance is determined, and mapped onto abstract lexical (word) representations that specify semantic and syntactic information. However, actual syntactic and sound planning occurs at a later stage of utterance planning. Ferreira et al. argue that ambiguity detection at utterance planning requires the speaker to run an internal comprehension-based monitoring process in parallel to the production system, a process that has independently been argued to be needed for the detection and repair of speech errors (see Postma for an overview). Because the time lag between the initiation of planning and the output of this system can easily exceed the time between initiation of planning and beginning of pronunciation, there is often insufficient time to incorporate changes in time for fluent conversation.

What about the ease with which referential indeterminacy is avoided at the message level? It may simply be the case that because message-level processes happen early in the production process, there is a greater time lag between the initiation of message-level planning and pronunciation, therefore allowing for greater opportunity for the monitoring process to take effect. Or, it may simply be the case that decisions about content at this level don't involve monitoring for potential difficulty from a specific hearer's perspective, but rather, recruit a set of heuristics attuned to unique referential identification from a set of contextually available referents. In other words, how much active calculation of the hearer's expectations and communicative needs is worked out by the speaker? There is a fair bit of evidence for partner-specific effects in production. Even young children, for instance, have been shown to be able to tailor their descriptions to their hearer's perspective rather than their own in the course of fluent speech (Nadig and Sedivy). Speakers have also been shown to be sensitive to the conversational history shared with a hearer. For example, they often choose a more specific description than is required by the immediate referential context if the same referent has previously been referred to by that description by either the speaker or hearer. This phenomenon has been called a 'conceptual pact' in conversation, reflecting an implicit agreement shared by partners for continuity of linguistic description (Brennan and Clark). However, there is also some evidence that suggests that partner-specific calculations impose a significant processing burden, and are vulnerable under time pressures (for further discussion, see Brown and Dell; Horton and Keysar; Lockridge and Brennan).

The work described here is part of a still-emerging body of data. However, we are in a position to see some patterns and reach the following tentative conclusions:

1) Due to internal properties of the production system, avoidance of form-based ambiguity seems to be difficult to achieve consistently in running speech. Hence, this particular aspect of the maxim of manner, in which the speaker's processing concerns collide with the hearer's, may be implemented only in quite limited ways.[3]
2) In contrast, avoidance of referential indeterminancy seems to be easily achieved in conversational speech. This may reflect a greater time lag between the massage formulation stage and onset of pronunciation. In addition, some evidence suggests that mechanisms that ensure successful, unique reference operate with great speed. Speakers are therefore well able to avoid difficulties for establishing reference that would arise for the hearer due to insufficient quantity of information. These findings suggest that there may be a great deal of variability in the reliability with which speakers can implement the various maxims under time pressure. This variability is likely to systematically reflect the internal workings of the processing system. Hence, a systematic study of the information processing capabilities of speakers is needed in order to have a detailed, realistic picture of the nature of the exchanges upon which speakers and hearers based their shared presumptions.
3) The message formulation component of the processing system shows considerable sensitivity to the hearer's perspective or communicative needs, suggesting that speakers do indeed calculate the potential consequences of their utterance for the hearer. However, these

partner-specific effects exert a discernible computational cost, and may at times 'run behind' the production system as utterance planning occurs.

4) There are no categorical constraints on whether a particular conversational maxim will be implemented during production. Rather, adherence to the maxims reflects the interaction of their computational requirements with the temporal constraints on processing at the particular stages at which these computations occur.

The situation-specific variability of adherence to conversational maxims raises a critical challenge for the hearer: given that conversational inferences deriving from the maxims rely on the hearer's expectations of what the speaker would say if adhering to the maxims, how well can the hearer adapt to what the speaker can actually be expected to say under duress of processing pressures? Can the hearer compute, for example, that when the complexity of an utterance increases, monitoring for potential ambiguity or for the hearer's perspective may degrade? If so, can the hearer flexibly suspend inferences that would arise out of an expectation of adherence to the maxims? We currently have few answers to these questions, and the field has only begun to explore the processing resources that are required for the interpretation of conversational implicature. In the next section, we review some of the first explorations in Gricean inferencing from the hearer's perspective.

3. Processing Issues for the Hearer

A starting point for investigation is simply to ask what the processing cost is for the hearer in computing a Gricean inference. Levinson's position anchors the debate at one end of the spectrum, with the suggestion that the speed of computing generalized implicatures is highly efficient and automatized, due to its computational parsimony. Therefore, we might expect generalized implicatures at least to be computed almost as efficiently as conventional meanings. However, some experimental work suggests a significant processing cost even for stereotypical quantity-based generalized implicatures. Bott and Noveck investigated processing times for making truth value judgments about the conventional versus inferentially enriched meanings of sentences such as 'Some elephants are mammals'. Note that this statement is true under the conventional meaning, but false under the enriched (*some but not all*) meaning. They found that response times were slower when the sentence was rejected as false (presumably based on the enriched meaning) than when it was accepted as true.[4] In addition, Bott and Noveck found that when pressured to respond within 9/10 of a second, participants were more likely to judge the test sentences to be false than when they were allowed to respond at leisure within a 3-second window, suggesting that under time pressure, they computed the conventional meaning only. Furthermore, when participants were coached to respond based on the conventional meanings, they responded faster than when they were coached to respond according to a pragmatically enriched interpretation of the test sentences. In the comprehension of running speech, this is quite relevant because a significant lag between the time to compute a conventional versus pragmatically enriched meaning may mean that the computation of the pragmatic enrichment may have to be abandoned in order to keep up with the incoming information in the speech stream. In the Bott and Noveck studies, the response times reflected the time to make a truth value judgment in addition to the time to interpret the sentence, therefore it is hard to draw concrete, quantitative conclusions about processing times of Gricean inferencing. Nevertheless, the work raises interesting implications for real-time comprehension.

Bott and Noveck argue that, in addition to demonstrating a measurable processing cost for quantity-based implicatures, their experiments provide evidence against the view that generalized conversational implicatures are generated automatically by default by the hearer, to be withdrawn only if there is a clash with the context. If this were the case, they claim, then the interpretations which reflect the sentence's conventional meanings (and hence, the withdrawal of an automatically generated implicature) should be taking longer than the interpretations in which the implicature is retained. They conclude that the data do not show support for the automatized nature of generalized implicatures.

Breheny, Katsos, and Williams make a similar point, using an experimental procedure based on a more naturalistic reading of short narratives. They used the disjunctive operator *or* as a test case, and compared contexts which either supported, or clashed with the quantity-based implicature associated with it.[5] Example narratives are in (12) and (13) below:

(12) *Implicature supporting context*:
John was taking a university course and working at the same time. For the exams, he had to study from short and comprehensive sources. Depending on the course, he decided to read the class notes or the summary.

(13) *Implicature clashing context*:
John heard that the textbook for Geophysics was very advanced. Nobody understood it properly. He heard that if he wanted to pass the course, he should read the class notes or the summary.

Breheny et al. reasoned that if the implicature is generated by default, and cancelled in the event of a clash with the context, then the phrase containing 'or' should be read faster in narratives such as (12), where the implicature goes through, than in (13), where it does not. On the other hand, if the implicature is built from the ground up, taking into account the context, just as particularized implicatures are assumed to be, then reading times should be longer in (12), which requires additional inferential work to generate the implicature, than in (13), in which no implicature is computed. Their results favored the latter account, showing longer reading times for the phrase containing *or* when it occurred with implicature supporting contexts.

Breheny et al. reported additional experiments which indicated that quantity-based implicatures associated with quantifiers such as 'some' may not be invariably computed, but depend on factors such as sentence position and preceding narrative context. Consider for example, the following narratives:

(14) Mary asked John whether he intended to host all his relatives in his tiny apartment. John replied that he intended to host some of his relatives. The rest would stay in a nearby hotel.

(15) Mary was surprised to see John cleaning his apartment and she asked the reason why. John told her that he intended to host some of his relatives. The rest would stay in a nearby hotel.

In the event that the *some but not all* reading had been computed, there should be no difficulty in reading the phrase 'the rest', which makes reference to the complement set of relatives. The results show longer reading times of this phrase in contexts such as (15) than (14). Furthermore, the phrase containing the triggering quantifier 'some' took longer to read in (14) than in (15). Breheny et al. interpret these results as suggesting that in contexts such as (14), the implicature was computed upon encountering 'some', requiring some additional processing effort, but subsequently making reference to the complement set easier to integrate than when the implicature was not initially generated upon encountering the quantifier. Similar results were found for sentence pairs in which the position of the quantified phrase was manipulated: when the quantifier phrase occurred in subject position, this phrase took longer to read relative to a control phrase than when it occurred in sentence final position; conversely, subsequent reference to a complement set took less time to read when the quantifier occurred in initial position, suggesting that the likelihood of computing a quantity-based implicature depends in part on structural factors, and not merely the presence of the triggering expression. Breheny et al. conclude that because of the context-dependence of these inferences, they do not have the properties associated with generalized conversational implicature.

However, while these studies show that *whether* these implicatures are computed depends on contextual and structural manipulations, I believe it would be premature to conclude that this is evidence against the generality of the *mechanism* by which such implicatures are generated, and an argument that they must reflect particularized inferencing. An alternative interpretation is that the effect of the context is to enhance (or reduce) the cognitive accessibility of the alternative expressions that the speaker might have used but did not. That is, the computations used to arrive at the inference may be general, but the accessibility of alternative expressions that serve as the input to this computation may vary according to context. Further research would be required to distinguish between these two quite different interpretations.

In the previous section, we saw that there are observable limits on a speaker's capacity to avoid ambiguity that is potentially costly to the hearer. However, we also saw that speakers are impressively good at avoiding referential indeterminacy, and generally provide as much information as is needed to establish unique reference. Mirroring these questions, we might ask whether hearers generate expectations that speakers will adhere to the quantity maxim in their choice of referring expression, and if so, whether such expectations are quickly integrated into the processing of an utterance. An interesting test case revolves around the interpretation of temporary ambiguities such as the example we discussed in (3) above, reproduced below with its possible continuation structures (4) and (5).

(3) The student examined . . .
(4) . . . the test tube before making notes.
(5) . . . by the committee was awarded her doctorate.

The syntactic ambiguity of 'the student examined . . .' introduces a contrast between a simple referring expression with no modifiers ('the student'), as in structure (4), and a complex referring expression with modification as in (5) ('The student examined by the committee'). Importantly, the two referring expressions differ in terms of the quantity of information they communicate. It turns out that quite a variety of temporary syntactic ambiguities involve just this contrast. Experimental evidence has shown that addressees typically find it hard to process structures which contain a complex, modified referring expression. While some researchers have argued that this reflects a preference for certain kinds of syntactic *structures* over others (e.g. Frazier), others have claimed that these effects turn directly on expectations about the amount of information provided in referring expressions (Crain and Steedman). That is, in the absence of a context in which a modifier phrase is required to distinguish between two entities denoted by the head noun, people will have a default expectation for a simple, unmodified description, and will have difficulty when this expectation is violated. Evidence for this latter view comes from a number of experiments showing that the preference for structures containing the simple referring expression (as in 4) is reduced or eliminated in contexts in which the modifier is in fact required for unique identification of the referent (Altmann and Steedman; Tanenhaus et al.). The following examples from Altmann and Steedman illustrate this effect. Here, the temporary ambiguity in question involves the location of the attachment of a prepositional phrase introduced by 'with'. One attachment results in a simple referring expression, with the prepositional phrase modifying the verb, and the other results in a complex referring expression:

(16) *Example context supporting complex referring expression*:
A burglar broke into a bank carrying some dynamite. He planned to blow open a safe. Once inside, he saw that there was a safe with a new lock and a safe with an old lock.

(17) *Target sentence resulting in a simple referring expression*:
The burglar blew open the safe with the dynamite and made off with the loot.

(18) *Target sentence resulting in a complex referring expression*:
The burglar blew open the safe with the new lock and made off with the loot.

Following contexts such as (16), target sentences such as (17) which contained a simple referring expression (e.g.'the safe') were harder to read than sentences such as (18) which contained the complex referring expressions (e.g.'the safe with the new lock'.) Exactly the reverse was true for the context illustrated in (19) below, which provided no communicative motivation to use the more informative referring expression:

(19) A burglar broke into a bank carrying some dynamite. He planned to blow open a safe. Once inside, he saw that there was a safe with a new lock and a strongbox with an old lock.

This line of argument suggests that hearers (or readers) are able to very rapidly consult expectations pertaining to the degree of informativeness in referring expressions, quickly enough to impact processing decisions about linguistic structure in real-time.[6]

This claim has been considered quite controversial in the processing literature, precisely because of commonly held assumptions that the kind of inferential processing that is required to generate these referential expectations is slower than the processing of structural information and its associated conventional meaning, and indeed, must depend on first having computed the conventional meaning. This objection is articulated by Clifton and Ferreira who conclude that for these reasons, conversational implicatures 'could not reasonably affect the initial steps of parsing'. However, there is further experimental evidence arguing for very rapid integration of informativeness expectations. Sedivy et al. conducted a study measuring eye movements to visual displays in response to spoken referring descriptions containing a prenominal adjective (e.g.'Pick up the tall glass'). The displays were such that upon hearing the adjective, more than one referent was possible (e.g. there were two objects that could be described as tall, such as the target class and a pitcher), thereby creating a referential indeterminacy at the adjective itself. Consistent with previous experimental work, this study showed that upon hearing the adjective, people typically looked at either of the two objects that matched its semantic content (e.g. the target glass and a pitcher). However, what was interesting was that a bias could be induced for one object simply by including in the display a contrasting object of the same kind, so that additional linguistic information would be needed to be distinguish between them. For instance, if the display contained a short glass in addition to the target glass and the pitcher, upon hearing 'tall', hearers were faster to look at the glass and less likely to look at the pitcher than when neither of the tall objects were contrasted with another object of the same kind. Thus, when the display contained a referent for which the use of a modifier was communicatively motivated, people showed a preference for this referent compared to displays in which there was no clear reason to refer to the same target referent using a modifier.

Additional experiments have shown that this effect generalizes beyond scalar adjectives, and seems to be very tightly linked to the actual tendencies of speakers to use modification in referring tasks. The degree to which the hearer infers a contrastive function has been shown to be linked to patterns of production for a number of different semantic classes of prenominal modifiers (Sedivy). For example, under some circumstances, speakers readily include color adjectives in a referential task even when these are not required to establish unique reference. That is, they might say 'Pick up the red bowl' even when there is a single bowl in the display.[7] Under the circumstances, where speakers are highly likely to include the modifier as part of their default description of the object as in 'the red bowl', hearers do not show a bias in their eye movement patterns for interpreting 'red' as singling out a member of a contrasting pair. However, when the color is readily inferrable from the object's category, speakers tend not to produce the color adjective except when needed for distinguishing the referent from another (that is, they would be unlikely to say 'yellow banana' when there is just one banana). When the modifier is not a part of the object's default description, its presence triggers the inference of contrastive function. This is intriguing, because it suggests that hearers have quite fine-grained expectations about what content speakers are likely to include in referring expressions, and use these expectations as a basis for ascertaining the communicative function of additional information. Furthermore, this appears to be achieved rapidly enough to be reflected in the course of making decisions about the denotation of predicates as they are being heard.

Do such effects in processing engage inferences of speaker rationality, or do they reflect mechanisms that reflect heuristics linking linguistic expressions fairly directly to their inferred function? Grodner and Sedivy found that hearers suspended the typical contrastive inference triggered by adjectives when faced with evidence of an uncooperative speaker. In this study, half of the hearers were told that they would be listening to recordings produced by a speaker who suffered from a neurological impairment leading to linguistic and social deficits, and heard a recording which contained a high degree of redundancy in the referring descriptions. These hearers did not show evidence of a contrastive bias in their eye movement patterns, unlike the other half of the hearers, who were presented with a 'normal' speaker. This indicates an ability to fine-tune expectations based on information about the speaker's likely adherence to conversational maxims and confirms the pragmatic, rather than conventional, nature of this aspect of interpretation. It also suggests that quantity-based expectations pertaining to reference are not automatically computed in response to the triggering expressions (though when they are, they can be computed very quickly in at least some circumstances, as shown by the eye movement data). That is, the eye movement record does not show evidence for a stage in processing in which the contrastive inference is first considered, and then later cancelled. Rather, the impact of the hearer's assessment of the speaker's reliability in adhering to the quantity maxim is evident in the earliest moments of reference resolution.

While systematic cognitive studies of pragmatic processes are still in their very early stages, based on the current work, we may arrive at the following tentative conclusions regarding the real-time interpretation of implicature:

(1) There is no evidence for the automatic generation of quantity-based implicatures in which context-specific cancellation occurs at a later processing stage. Rather, implicatures associated with the maxim of quantity show some degree of context-sensitivity.
(2) We do not yet have a clear picture of the hearer's processing costs of computing implicatures; while some studies show a measurable processing cost, others suggest that Gricean inferencing can be used without detectable cost, and efficiently enough to constrain the referential interpretation of incoming linguistic material.
(3) Certain very rapid, incremental inferences deriving from quantity-based expectations show sensitivity to evidence about the speaker's likely adherence to Gricean maxims, suggesting that even the most efficient Gricean inference cannot be undertaken without consideration of speaker meaning.

Acknowledgments

This work has been funded in part by NIH grant R01 MH62566. I am grateful to Shaun Nicols and an anonymous reviewer for *Philosophy Compass* for comments on an earlier version of this article.

Short Biography

Julie Sedivy's research uses experimental techniques to address cognitive questions about semantics and pragmatics, particularly as they relate to issues of real-time language processing. She is especially interested in the interplay between conventional linguistic knowledge and contextual information and expectations. Her current work focuses largely on pragmatic inferencing during real-time comprehension, with the aim of characterizing the cognitive mechanisms and consequences of conversational implicature. She has authored and co-authored articles investigating contextual influences on language processing in journals such as *Science*, *Cognition*, *Journal of Memory and Language*, and *Cognitive Psychology*, and is currently the associate editor for Psycholinguistics for the journal *Linguistics and Philosophy*. As Associate Professor in the Department of Cognitive and Linguistic Sciences at

Brown University, she teaches graduate and undergraduate courses in psycholinguistics and semantics/pragmatics. She holds a B.A. in Linguistics from Carleton University, a M.A. in Linguistics from the University of Ottawa, and a Ph.D. in Linguistics from the University of Rochester.

Notes

* Correspondence address: Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912, USA. Email: Julie_Sedivy@brown.edu.

[1] For an overview of the problem of ambiguity resolution for comprehension, see Tanenhaus and Trueswell; Frazier.

[2] I use the term '*referential indeterminacy*' here to describe cases where a single conventional meaning maps onto multiple possible referents and, to distinguish these from cases of *linguistic ambiguity*, where a single word or string of words maps onto multiple possible conventional meanings.

[3] However, other aspects of the manner maxim seem well-aligned with production considerations, and I suspect, are likely to be robustly and reliably implemented. For example, avoidance of unusual or prolix forms falls naturally out of the pressures on the production system in that more frequent or usual forms will be more readily available to the processing system. Similarly, the tendency to generate orderly narratives may reflect natural message formulation processes, though Carston ('Relevance Theory') has pointed out some interesting situations in which the processing needs of the speaker and hearer may be misaligned in this regard.

[4] An additional comparison within the experiment also determined that rejecting the enriched meaning was slower than rejecting other control sentences which were false on the basis of their conventional meanings, suggesting that the increased processing time was not simply due to the fact that judging a sentence as false takes longer than judging one to be true.

[5] By the quantity maxim, a speaker making a statement such as 'Walt is intelligent or ambitious' would typically implicate that Walt was not both intelligent and ambitious, because had he intended to communicate the latter thought, he would have chosen the stronger (more informative) connective '*and*' rather than '*or*'.

[6] Researchers originally characterized these effects as deriving from linguistic presupposition rather than implication (Crain and Steedman; Altmann and Steedman). However, along with Clifton and Ferreira, I have argued elsewhere ('Pragmatic versus Form-Based Accounts of Referential Contrast') that the referential expectations reflect implicature, and not presupposition).

[7] This tendency to produce seemingly redundant information has been noted in several studies, and may reflect very high salience and accessibility of the 'extra' linguistic expressions (Engelhardt, Bailey, and Ferreira) or it may serve some function in orienting the hearer to a salient property of the referent (Deutsch and Pechmann).

Works Cited

Allbritton, D. W., G. McKoon, and R. Ratcliff. 'Reliability of Prosodic Cues for Resolving Syntactic Ambiguity'. *Journal of Experimental Psychology: Learning, Memory & Cognition* 22 (1996): 714–35.

Altmann, G., and M. Steedman. 'Interaction with Context during Human Sentence Processing'. *Cognition* 30 (1988): 191–238.

Arnold, J. E., T. Wasow, A. Asudeh, and P. Alrenga. 'Avoiding Attachment Ambiguities: The Role of Constituent Ordering'. *Journal of Memory and Language* 51 (2004): 55–70.

Bach, K. 'The Top 10 Misconceptions about Implicature'. *Drawing the Boundaries of Meaning: Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn. Studies in Language Companion Series, Volume 80.* Eds. B. J. Birner and G. Ward. Amsterdam: John Benjamins, 2006. 21–30.

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

Bott, L. and I. A. Noveck. 'Some Utterances are Underinformative: The Onset and Time Course of Scalar Inferences'. *Journal of Memory and Language* 51 (2004): 437–57.

Breheny, R., N. Katsos, and J. Williams. 'Are Generalized Scalar Implicatures Generated by Default? An On-Line Investigation into the Role of Context in Generating Pragmatic Inferences'. *Cognition* 100 (2006): 434–63.

Brennan, S. E. and H. H. Clark. 'Conceptual Pacts and Lexical Choice in Conversation'. *Journal of Experimental Psychology: Learning, Memory & Cognition* 22 (1996): 1482–93.

Brown, P. M. and G. S. Dell. 'Adapting Production to Comprehension: The Explicit Mention of Instruments'. *Cognitive Psychology* 19 (1987): 441–72.

Brown-Schmidt, S. and M. K. Tanenhaus. 'Watching the Eyes When Talking about Size: An Investigation of Message Formulation and Utterance Planning'. *Journal of Memory and Language* 54 (2006): 592–609.

Carston, R. 'Informativeness, Relevance and Scalar Implicature'. *Relevance Theory: Applications and Implications*. Eds. R. Carston and S. Uchida. Amsterdam: John Benjamins, 1998. 179–236.

——. 'Linguistic Meaning, Communicated Meaning, and Cognitive Pragmatics'. *Mind & Language* 17 (2002): 127–48.

——. 'Relevance Theory, Grice and the Neo-Griceans'. *Intercultural Pragmatics* 2.3 (2005): 303–19.

Chierchia, G. 'Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface'. *Structures and Beyond*. Ed. A. Belletti. Oxford: Oxford UP, 2004. 39–103.

Clifton, C. and F. Ferreira. 'Ambiguity in Context'. *Language and Cognitive Processes* 4 (1989): 77–103.

Crain, S. and M. Steedman. 'On Not Being Led up the Garden Path: the Use of Context by the Psychological Parser'. *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. Eds D. Dowty, L. Karttunen and A. Zwicky. Cambridge, MA: Cambridge University Press, 1985. 320–58.

Davis, W. *Implicature: Intention, Convention and Principle in the Failure of Gricean Theory*. Cambridge: Cambridge UP, 1998.

Deutsch, W. and T. Pechmann. 'Social Interaction and the Development of Definite Descriptions'. *Cognition* 11 (1982): 159–84.

Elsness, J. 'That or Zero? A Look at the Choice of Object Clause Connective in a Corpus of American English'. *English Studies* 65 (1984): 519–33.

Engelhardt, P. E., K. G. D. Bailey, and F. Ferreira. 'Do Speakers and Listeners Observe the Gricean Maxim of Quantity?'. *Journal of Memory and Language* 54 (2006): 554–73.

Ferreira, V. and G. S. Dell. 'Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production'. *Cognitive Psychology* 40 (2000): 296–340.

Ferreira, V. S., L. R. Slevc, and E. Rogers. 'How do Speakers avoid Ambiguous Linguistic Expressions?'. *Cognition* 96 (2005): 263–84.

Frazier, L. 'Theories of Sentence Processing'. *Modularity in Knowledge Representation and Language Processing*. Ed. J. Garfield. Cambridge, MA: MIT Press, 1987. 291–308.

Grice, H. P. 'Logic and Conversation'. *Syntax and Semantics, Vol 3, Speech Acts*. Eds. P. Cole and J. Morgan. New York, NY: Academic Press, 1975. 43–58.

Grodner, D. and J. Sedivy. 'The Effect of Speaker-Specific Information on Pragmatic Inferences'. *The Processing and Acquisition of Reference*. Eds. N. Pearlmutter and E. Gibson. Cambridge, MA: MIT Press, forthcoming.

Horn, L. R. 'Implicature'. *The Handbook of Pragmatics*. Eds. L. R. Horn and G. Ward. Malden, MA: Blackwell Publishers, 2004. 3–28.

——. 'Toward a New Taxonomy for Pragmatic Inference: Q- and R-Based Implicature'. *Meaning, Form and Use in Context*. Ed. D. Schiffrin. Washington, DC: Georgetown UP, 1984. 11–42.

Horton, W. S. and B. Keysar. 'When Do Speakers Take into Account Common Ground?'. *Cognition* 59 (1996): 91–117.

Kraljic, T. and S. E. Brennan. 'Prosodic Disambiguation of Syntactic Structure: For the Speaker or for the Addressee?'. *Cognitive Psychology* 50 (2005): 194–231.

Levelt, W. J. M. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press, 1989.

Levinson, S. *Presumptive Meanings*. Cambridge, MA: MIT Press, 2000.

Lockridge, C. B. and S. E. Brennan. 'Addressees' Needs Influence Speakers' Early Syntactic Choices'. *Psychological Bulletin & Review* 9 (2002): 550–7.

Matsumoto, Y. 'The Conversational Condition on Horn Scales'. *Linguistics and Philosophy* 18 (1995): 21–60.

Metzing, C. and S. E. Brennan. 'When Conceptual Pacts are Broken: Partner-Specific Effects in the Comprehension of Referring Expressions'. *Journal of Memory and Language* 49 (2003): 201–13.

Nadig, A. S. and J. C. Sedivy. 'Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution'. *Psychological Science* 13 (2002): 329–36.

Olson, D. R. 'Language and Thought: Aspects of a Cognitive Theory of Semantics'. *Psychological Review* 101 (1970): 676–703.

Postma, A. Detection of Errors during Speech Production: A Review of Speech Monitoring Models. *Cognition* 77 (2000): 97–131.

Potts, C. 'Conversational Implicatures via General Pragmatic Pressures'. *Preproceedings of Logic Engineering and Natural Language Semantics*. Ed. E. McCready. Tokyo: Japanese Society for Artificial Intelligence, 2006. 65–79.

Saul, J. 'What is Said and Psychological Reality: Grice's Project and Relevance Theorists' Criticisms'. *Linguistics and Philosophy* 25 (2002): 347–72.

Schafer, A. J., S. Speer, P. Warren, and S. D. White. 'Intonational Disambiguation in Sentence Production and Comprehension'. *Journal of Psycholinguistic Research* 29 (2000): 169–82.

Sedivy, J. 'Pragmatic versus Form-Based Accounts of Referential Contrast: Evidence for Effects of Informativity Expectations'. *Journal of Psycholinguistic Research* 32 (2003): 3–23.

——, M. Tanenhaus, C. Chambers, and G. Carlson. 'Achieving Incremental Semantic Interpretation through Contextual Representation'. *Cognition* 71 (1999): 109–47.

Snedeker, J. and J. Trueswell. 'Using Prosody to avoid Ambiguity: Effects of Speaker Awareness and Referential Context'. *Journal of Memory and Language* 48 (2003): 103–30.

Sperber, D. and D. Wilson. *Relevance: Communication and Cognition*. Oxford: Blackwell, 1986.

Tanenhaus, M. K., M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. 'Integration of Visual and Linguistic Information in Spoken Language Comprehension'. *Science* 268 (1995): 1632–4.

—— and J. C. Trueswell. 'Sentence Comprehension'. *The Handbook of Perception and Cognition*. Vol. 11. Eds. J. Milller and P. Eimas. New York, NY: Academic Press, 1995. 217–62.