Semantic Ambiguity Resolution as a Decision Process

Nicholas Gaylord (nlgaylord@utexas.edu)
Colin Bannard (bannard@utexas.edu)
Department of Linguistics, University of Texas at Austin
305 E. 23rd St. B5100, Austin, TX 78751 USA

Abstract

Resolution of the meaning of a semantically ambiguous word requires knowledge about the space of possible meanings of that word, and the selection of a meaning in the light of available evidence and given situational constraints. As such, ambiguity resolution bears many similarities to decision making scenarios more generally. We report on an experiment exploring this analogy by applying some standard manipulations from the decision making literature to a semantic disambiguation task. We explore two particular proposals: (1) that depth of semantic processing can be cast as strategy selection reflecting a risk-sensitive effort-accuracy tradeoff, and (2) that thresholds for inference about meaning in context are situationally flexible and learnable via feedback. One robust property of decision making is people's ability to use feedback in order to adjust responses to maximize payoffs. Participants completed a semantic entailment judgment task in which they received trial-by-trial feedback, and payoff matrices and decision thresholds were manipulated across conditions. We find an effect of risk, with participants employing different comprehension strategies depending on relative gains and losses. We also find that participants were in fact sensitive to varying decision thresholds and accurately adjusted their behavior to match the constraints on what qualified as a true conclusion in different conditions. We take these findings as preliminary evidence that ambiguity resolution in language can be modeled, at least in part, as involving more general decision processes.

Keywords: Speed-Accuracy Tradeoff, risk, decision thresholds, decision making, sentence processing, word meaning, ambiguity

Introduction

Semantic ambiguity is a widespread phenomenon in natural language, whereby a single word can have more than one interpretation depending on its use. The resolution of semantic ambiguity requires knowledge of the range of possible meanings of an ambiguous word, and the consideration of those possibilities in light of available contextual evidence and given certain situational constraints (such as how strict or precise an interpretation is required). Characterized as such, semantic ambiguity resolution bears many similarities to other scenarios that are studied in research on human decision making. However, the connection between decision making and semantic processing is as yet underexplored.

One robust property of human decision making in other domains is the ability to use feedback in order to adjust responses to maximize benefits (increasing material rewards and/or minimizing cognitive costs). In this paper we look at whether the same behavior might be observed for a semantic disambiguation task. Two particular manipulations were employed, parallel to manipulations in other decision tasks: a) changes to the decision threshold, which separated correct "true" or "false" responses concerning the meaning of a word in context, and b) changes to the degree of risk (possible material losses) in the decision situation. Such factors have useful analogues in language understanding. Decision threshold changes are implicated in that different situations call for more (or less) restrictive assumptions as to what can be safely concluded from a potentially ambiguous utterance. Risk is implicated via the potential negative consequences of misinterpretation, which is greater in some cases than others – for example, a failure of interpretation is likely more consequential in a job interview than in a casual conversation.

Background

Semantic ambiguity has been extensively studied from a variety of perspectives including linguistic theory and psycholinguistics. One important finding from this work is that not all cases of semantic ambiguity are the same – Apresjan (1974) argues that different senses (or uses) of a word can vary in how semantically similar they are, and most psycholinguistic research into the representation of semantic ambiguity arrives at a similar conclusion (Brown, 2008; Frazier & Rayner, 1990; Klepousniotou, Titone, & Romero, 2008; Pickering & Frisson, 2001; Williams, 1992). This position is further supported by various offline judgment tasks (Erk, McCarthy & Gaylord 2009, To Appear; Gaylord, 2011). In short, there is a growing body of evidence that word meanings are graded – the meanings of individual occurrences of a word can vary quite subtly, and the extent to which word senses apply to a given occurrence varies in a graded fashion as well.

A closely related question is that of how we use the information available in our lexical representations to determine a contextually-appropriate meaning. McElree, Murphy, and Ochoa (2006) and Gaylord, Goldwater, Bannard, and Erk (2012) both investigated the dynamics of this process using a Speed-Accuracy Tradeoff (SAT) design. McElree, Murphy, and Ochoa observed elevated false alarms after short processing delays with stimuli such as *Water pistols – are dangerous* and Gaylord *et al.* found the same effect with stimuli such as *The dawn broke – Something shattered*. In other words, both studies found evidence that when a word is encountered, a context-independent default meaning is activated prior to semantic integration, whether or not it is supported by the occurrence in question. It is likely that these default meanings correspond to those words' most frequent interpretations. There is a current debate as to how information-rich our lexical knowledge must be (cf. Elman 2011) and while evidence is accumulating that our semantic representations provide access to a great deal of richly informative world knowledge, results such as those just discussed also indicate that our knowledge of word meanings contains a more schematic layer that is more efficient to access.

One question that can be raised is why this more schematic level of semantic representation is present despite the fact that it can lead to errors of interpretation. A plausible answer is that sentence comprehension, like other cognitive processes, is subject to economic pressures, and that under many circumstances a shallower processing is sufficient (Barton & Sanford, 1993; Bever, Sanz, & Townsend, 1998; Ferreira & Patson, 2007; Sturt, Sanford, Stewart, & E Dawydiak, 2004; Swets, Desmet, Jr., & Ferreira, 2008; Townsend & Bever, 2001). We hypothesize that these default meanings, as they reflect a word's most likely interpretation, support a shallow semantic processing strategy.

A considerable amount of decision making research addresses the question of strategy selection (Beach & Mitchell, 1978; Busemeyer, 1993; Gigerenzer, Todd, & ABC Research Group, 1999; Johnson & Payne, 1985; Payne, Bettman, & Johnson, 1988). This work studies how people select more or less effortful decision making strategies in different situations, where increased effort tends to yield increased accuracy. The concept of an effort-accuracy tradeoff is central to strategy selection, and has been seen to be sensitive to risk. Semantic comprehension has been shown to be effortful, and we propose that depth of semantic processing can be cast as a strategy selection problem driven by a risk-sensitive effort-accuracy tradeoff. We explore this hypothesis through changes across conditions to the payoff matrix dictating potential gains and losses for correct and incorrect responses. We hypothesize that shallow processing strategies (marked by acceptance of default meanings in the absence of contextual support) will be more prevalent under decreased risk, and dispreferred when potential losses are high.

However, parallels with strategy selection are not the only similarity between semantic comprehension and decision making more generally. More generally, ambiguity resolution requires the selection of a possible interpretation of a word in light of available contextual evidence, and given situational constraints on interpretation. As discussed above, meaning-in-context appears to be a very graded phenomenon, and another question is whether people adapt their semantic comprehension behavior to meet situational demands. We explore this question as well by moving the threshold (corresponding to a property of the stimulus) at and above which a response of "true" will be counted as correct in different conditions, and providing trial-by-trial feedback on response accuracy. We hypothesize that participants will use their graded representations of word meanings in order to rapidly learn an optimal decision threshold.

Experiment

Participants 131 undergraduate psychology students from the University of Texas at Austin completed the experiment in exchange for course credit. Participants received a cash payment of up to $3.00 depending on their performance on the task. All participants were native English speakers.

Table 1: Example stimuli, with their associated truth norms. TS = true given context sentence; PS = plausible given context sentence; FS = false (but possible given a different sentence); FV = false given the verb (false regardless of context)

Context	Probe	Norm	Type
The insult burned	Something was mean	6.65	TS
The insult burned	Something was true	4.85	PS
The insult burned	Something was warm	1.30	FS
The insult burned	Something was rolled	1.30	FV
The log burned	Something was warm	6.55	TS
The log burned	Something was dangerous	4.55	PS
The log burned	Something was mean	1.10	FS
The log burned	Something was fixed	1.20	FV

Table 2: Summary of experimental conditions. Threshold is the truth norm at and above which items were counted as true.

Condition	Gain/Loss	Threshold
A	+5 / -1	3.7
B	+1 / -5	3.7
C	+5 / -5	3.7
D	+5 / -5	2
E	+5 / -5	6

Materials The experiment, 240 trials in length, took the form of a semantic judgment task in which each trial consisted of a context sentence (e.g. *The dawn broke*) followed by a semantic probe (e.g. *Something shattered*) to be evaluated as true or false. Each context sentence was paired with a true probe, a plausible but not necessarily true probe, and two false probes: one which would be true under a different meaning of the context verb, and one which was false given any contextually-activated meaning of the verb. The truthfulness of each probe, given its context sentence, was measured via a separate offline norming task and the averages of these ratings established a truthfulness value for each stimulus. Further examples of stimuli are contained in Table 1.

Procedure Participants were told that the experiment would take the form of a game, in which points were gained or lost based on accuracy, and that those points were redeemable for cash at the end of the session. At the start of the experiment, participants were familiarized with the gains and losses associated with correct and incorrect responses, and after each trial they received feedback about their response accuracy (a smiling face for a correct response or a frowning face for an error) and associated gain or loss of points.

A schematic of an experimental trial is shown in Figure 1. The experiment contained 5 conditions (summarized in Table 2), across which we varied risk (via changes to payoff matrices, which were always symmetrical) and decision threshold.

Risk. In condition A participants could gain 5 points for a correct answer, but only lose 1 point for an incorrect answer.

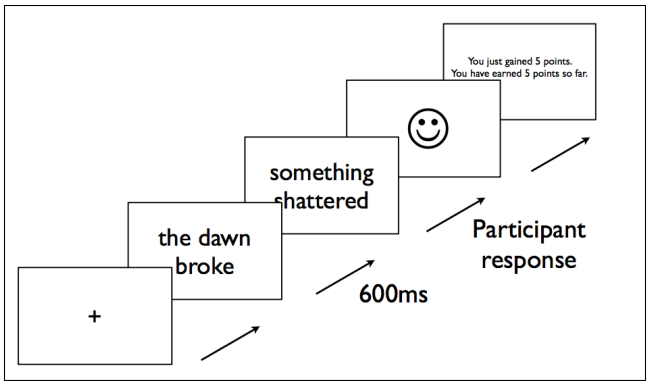

Figure 1: A single trial, depicting a correct "false" response.

Hence the risk associated with giving an errorful response was minimal. In condition B, participants could gain only 1 point for a correct answer but could lose 5 points for an incorrect answer. In condition C participants could gain 5 points for a correct answer and lose 5 points for an incorrect answer. Both conditions B and C are characterized by large potential losses, and are by extension higher in risk than condition A. While the difference between possible gain and loss on a trial is smaller in condition B than in condition C, and is in fact exactly the same as in condition A, we know that people tend to evaluate risk relative to a status quo reference point and tend to be loss-averse (Kahneman & Tversky, 1979). As such, condition B is the highest risk case overall, since possible gains are much smaller than possible losses.

Acceptance Threshold. Across conditions C-E we varied the acceptance threshold while holding risk constant. Acceptance threshold manipulations were accomplished relative to the mean truthfulness ratings that we independently gathered. Very high-rated stimuli were true across conditions, and very low-rated stimuli were consistently false, but stimuli with intermediate truthfulness ratings were counted as true in some conditions but false in others. The threshold for condition C was at median. The threshold for conditions D and E were lower and higher respectively. These thresholds determined the feedback we gave to participants on their responses.

Results and Discussion

Participants' responses and response times were recorded on each trial. Participant responses of under 150 ms were excluded, as well as the 0.5% of slowest responses.

Effects of Risk. Risk was manipulated via changes to the study payoff matrix, such that in Condition A possible gains on each trial were large and possible losses small, while in Conditions B and C possible losses were greater. We hypothesized that participants would employ different response strategies in the higher-risk Conditions C (in which uninformed responding would yield a loss relative to the maximum points possible) and B (in which uninformed responding would be expected to yield a loss relative to the starting point), relative to the low-risk Condition A (in which uninformed responding would be expected to yield a net gain).

We performed a series of multilevel logistic regression models, in which the outcome was the participants' response (true=1, false=0) and in which participant ID was included as a random effect on the model intercept. We first of all examined simple accuracy by looking at whether correct response was a good predictor of actual participant responses across the conditions. As discussed in Wright and London (2009) this is equivalent to a traditional d-prime analysis. A model containing an interaction between correct response and condition was found to give a significantly better fit to the data than a model containing only correct response ($\chi^2(4) = 15.673$, $p < 0.01$) and a model containing both terms but no interaction ($\chi^2(2) = 11.172$, $p < 0.01$). The coefficients revealed the increase in the likelihood of participants responding "true" if the correct response was "true" was significantly greater in both Conditions B and C than it was in condition A.

We next looked in more detail at how participants were making their decisions. In our norming study we obtained graded ratings as to whether the probe sentences were entailed by the context sentences. We assume that participants in our main study were able to utilize intuitions that corresponded to such scales. We first looked at whether our normed truth scale was predictive of response in a series of logistic regression models. A model including the truth norm rating as a predictor gave a better fit to the data than a model including the correct response as sole predictor. A model containing an interaction between truth norm rating and condition was a significantly better fit than a model containing only truth norm ($\chi^2(4) = 17.794$, $p < 0.01$) or one containing both terms but no interaction ($\chi^2(2) = 13.228$, $p < 0.01$). The coefficients revealed the increase in the likelihood of participants responding "true" as a function of increases in the truth norm was significantly greater in both conditions B and C than in condition A. We next fit separate logistic regression models to the data from each of the conditions and looked at the predictive value of the truth norms in each case. Log Likelihood Ratio Indices (McFadden, 1974) revealed that the truth values had more predictive value in conditions B (0.458) and C (0.477) than in condition A (0.420). These data further support the finding that probe truthfulness is a stronger determinant of participant response under increased risk.

We performed a final exploration by defining a simple model based on these truthfulness values and exploring how well it accounts for participants' responses. We assume that an idealized responder would say "true" for a given item with a probability equal to the mean truth rating provided (minus the minimum possible response, 1), divided by the difference between the minimum and maximum response (6). We look at the perplexity (an information theoretic measure of how surprised the model is by the data) of such a model when confronted with participant response data. Model perplexity is higher for condition A (2.328) than it is for conditions

B (2.213) and C (2.110). This indicates that participant responses in the riskier conditions are better described by probe truthfulness than are responses in the less risky conditions.

These analyses suggest that participants are making more sensitive semantic judgements in the riskier conditions. Because semantic comprehension is effortful, one explanation of this is that there is an effort-accuracy tradeoff at work. Participants should be more willing to expend this effort via deeper semantic processing when there is more at stake. A possible consequence of this would be an increase in the time taken to make decisions. This effect is in fact seen, though response times are not elevated across-the-board in the riskier conditions. Rather, stimuli with very high or very low truthfulness values are processed as quickly as in the low-risk condition, and extra time is spent precisely those items that warrant it – items with intermediate truthfulness ratings.

Effects of Decision Threshold Placement. We next turn to the effect of changes to the decision threshold. In Conditions C–E, the same stimuli were used but the threshold value at and above which a probe was considered true was varied across conditions. Based on our above argument that word meaning in context is a graded phenomenon whose scales can be used flexibly in making decisions, we hypothesized that participants would adjust their responses to reflect these thresholds. Our norming study showed that people are able to reliably assign graded values as to whether our probe sentence was entailed by our context sentence. We assume that participants in our main experiment will have similar graded evaluations and that they will respond differently depending on our different conditions by inferring an optimal point on their scales at which to accept or reject probes. It is worth reiterating that this is not an arbitrary manipulation – different situations do indeed carry different constraints on meaning-in-context inference. A legal contract, for example, demands a very constrained interpretation of explicitly presented information, while innuendo demands much greater inference.

Acceptance probabilities (across participants) for all stimuli as a function of their truthfulness values are contained in Figure 2, in which it is clearly visible that participants do evaluate stimuli differently between conditions. This is particularly true for stimuli with intermediate (2–6) truthfulness ratings, which are evaluated differently depending on threshold placement. There is much less effect on the acceptance of very high- or very low-rated stimuli. The effect of threshold placement is further supported by the improved fit of a model with an interaction between item truth norm and condition, compared to a model with item truth norm as the only predictor ($\chi^2(4) = 191.69$, $p < 0.001$). Our primary interest, however, is in how rapidly participants learn different decision thresholds. Figure 3 sheds light on this question, showing the changes over trials in the minimum truthfulness ratings for the items that are accepted and the maximum ratings for the items that are rejected for the three conditions. This shows how participants adjust these cutoffs over trials differently in line with the acceptance thresholds revealed to them via feedback. This is supported by model comparison – a model including a three-way interaction between item truth norm, condition, and trial number gives a significantly better fit ($\chi^2(6) = 181.1$, $p < 0.001$). We take this as evidence that people do dynamically adjust their judgments about meaning-in-context, specifically how broadly or conservatively they interpret meaning, in response to situational constraints.

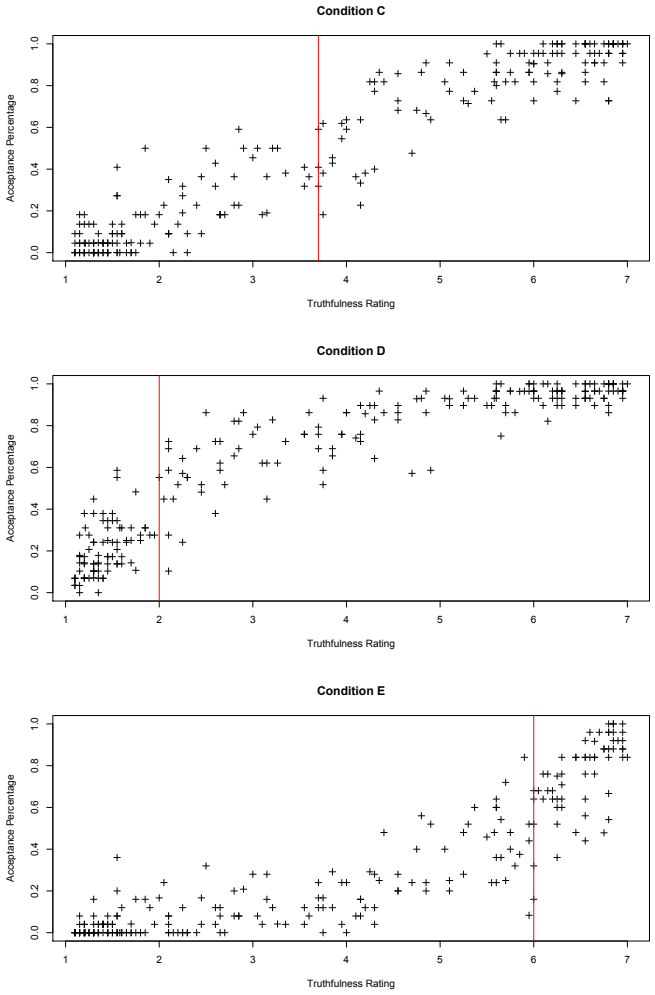

Figure 2: Acceptance probabilities for stimuli as a function of their truthfulness ratings in Conditions C–E. For each condition, the "true"/"false" threshold is indicated in red.

General Discussion

We found that participants dynamically adjust their assumptions regarding the conclusions that can be drawn from a given utterance in response to feedback. We also found that they employed different responding strategies depending on risk, and that the difference was not simply due to a speed-accuracy tradeoff. These results suggest that decision making behaviors that have been reported in other non-linguistic do-

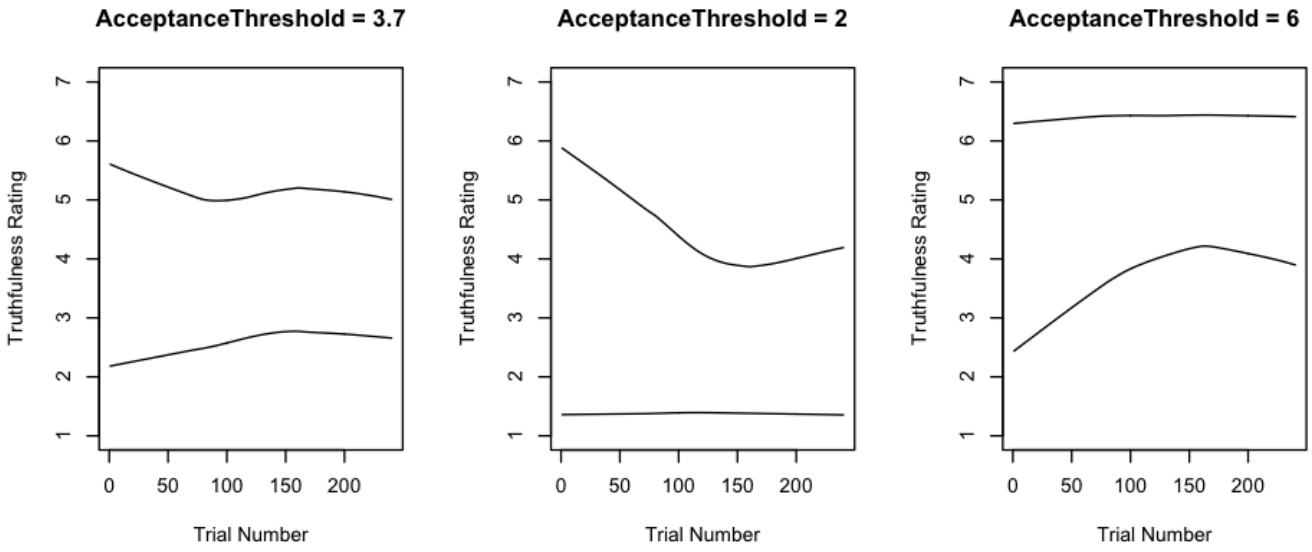

Figure 3: Lowess smoothed values for the minimum ratings at which items were accepted (the lower line) and maximum ratings at which items were rejected (the upper line) by trial for our conditions C-E.

mains might reasonably be extended to sentence processing, and in particular to meaning-in-context resolution.

The finding that semantic ambiguity resolution is affected by economic considerations is timely given related developments in the literature. One is the appearance in the sentence processing literature of the so-called "Good Enough" approach (Ferreira & Patson, 2007), along with other studies (discussed above) that have found people often only engage in shallow syntactic or semantic processing. To the best of our knowledge, the present work is the first to extend this line of inquiry to semantic processing at the lexical level, and the first study to explicitly predict semantic processing depth based on situational characteristics. The connection we make here with the decision making literature suggests further possibilities for studying the effect of situational pressures on language processing. Techniques from this literature, such as trial-by-trial feedback, are being adopted by other research in language processing as well (Lewis, Shvartsman, & Singh, To Appear) . Another recent development, this time in the theoretical linguistic literature, has been the use of ideas from decision theory and game theory to discuss linguistic communication (Clark, 2012) and particularly pragmatics (Benz, Jager, & Rooij, 2006). One of the main challenges in extending these accounts is the effective parameterization of utilities. Our findings suggest that standard techniques from the decision making literature might be useful in this regard.

We have argued here that participants' performance in the absence of risk reflects the use of default interpretations of the kind found by McElree et al. (2006) and Gaylord et al. (2012). Depending on the degree of situational risk, people might vary in how readily they will accept an initially-activated default interpretation of a word, presumably because under certain payoff schemes it is no longer worth the effort of computing a more precise interpretation to avoid a marginal potential loss. A related question to be explored in greater detail in future research is how this readiness to accept default interpretations is affected by the relative strength of the default meaning versus other potentially competing candidate interpretations (Kilgarriff, 2004), and in fact whether it is the case that only one default interpretation is activated.

Participants' dynamic adjustment to different truthfulness thresholds is equally striking as it shows that in different situations they rapidly learned how conservative or permissive to be regarding the possible conclusions that can be drawn from given information. This is particularly relevant for the study of semantic ambiguity resolution due to the fact that contextualized meaning has long been tied to the set of conclusions that can be drawn from a sentence. While approaches such as that in Chierchia and McConnell-Ginet (2000) are more restrictive in that they characterize sentence meaning through entailments, which are necessarily true, as opposed to here where we also deal with plausible conclusions, the general sentiment of these approaches is nonetheless applicable. Additionally, manipulations of decision threshold such as those employed here may prove useful to the broader study of inference in experimental pragmatics.

An immediate next step is to observe the effects of simultaneously varying both risk and decision threshold. We have already seen that participants rapidly learn to approximate the threshold, and we have seen that participants become more deliberative under higher risk. These facts jointly predict that threshold learning will be both more rapid and more accurate under increased risk. Investigation of these questions is currently underway.

Acknowledgments

This work was supported in part by a Carlota S. Smith Memorial Research Fellowship to the first author. We also wish to thank Arthur B. Markman for many very important insights as well as the use of his lab, and J. Grant Loomis for his invaluable assistance throughout the data collection process.

References

Apresjan, J. (1974). Regular polysemy. *Linguistics*, *142*, 5–32.

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, *21*, 477–487.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *The Academy of Management Review*, *3*, 439–449.

Benz, A., Jager, G., & Rooij, R. van. (2006). *Game theory and pragmatics*. New York: Palgrave Macmillan.

Bever, T. G., Sanz, M., & Townsend, D. J. (1998). The emperor's psycholinguistics. *Journal of Psycholinguistic Research*, *27*, 261–284.

Brown, S. (2008). Choosing sense distinctions for WSD: Psycholinguistic evidence. In *Proceedings of ACL-08: HLT, short papers (companion volume)* (p. 249-252). Association for Computational Linguistics.

Busemeyer, J. (1993). Violations of the speed-accuracy tradeoff relation. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making.* New York: Plenum Press.

Chierchia, G., & McConnell-Ginet, S. (2000). *Meaning and grammar: An introduction to semantics (2nd ed.)*. MIT Press.

Clark, R. (2012). *Meaningful games: Exploring language with game theory*. MIT Press.

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, *6*, 1–33.

Erk, K., McCarthy, D., & Gaylord, N. (2009). Investigations on word senses and word usages. In *Proceedings of ACL 2009.*

Erk, K., McCarthy, D., & Gaylord, N. (To Appear). Measuring word meaning in context.
(To appear in *Computational Linguistics*)

Ferreira, F., & Patson, N. D. (2007). The good enough approach to language comprehension. *Language and Linguistics Compass*, *1*, 71–83.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitements: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*, 181–200.

Gaylord, N. (2011). Exploring contextual effects on word meaning via multiple-level similarity judgments. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 950–955). Austin, TX: Cognitive Science Society.

Gaylord, N., Goldwater, M., Bannard, C., & Erk, K. (2012). Default verb meanings and verb meaning-in-context: A speed-accuracy tradeoff study. In *Proceedings of architectures and mechanisms for language processing (AMLaP 2012)* (p. 218). Riva del Garda, Italy.

Gigerenzer, G., Todd, P. M., & ABC Research Group the. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, *31*, 395–414.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? In Sojka, Kopecek, & Pala (Eds.), *Proceedings of Text, Speech, Dialogue. Lecture notes in artificial intelligence* (pp. 103–112). Springer.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1534–1543.

Lewis, R. L., Shvartsman, M., & Singh, S. (To Appear). The adaptive nature of eye-movements in linguistic tasks: How payoff and architecture shape speed-accuracy tradeoffs. *Topics in Cognitive Science*.

McElree, B., Murphy, G., & Ochoa, T. (2006). Time course of retrieving conceptual information: A speed-accuracy tradeoff study. *Psychonomic Bulletin and Review*, *13*, 848–853.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econonometrics* (pp. 105–142). New York: Academic Press.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 534–522.

Pickering, M., & Frisson, S. (2001). Processing ambiguous verbs: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 556–573.

Sturt, P., Sanford, A. J., Stewart, A., & E Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change- detection paradigm. *Psychonomic Bulletin & Review*, *11*, 882–888.

Swets, B., Desmet, T., Jr., C. C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*, 201–216.

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of rules and habits*. Cambridge, MA: MIT Press.

Williams, J. (1992). Processing polysemous words in context: Evidence for interrelated meanings. *Journal of Psycholinguistic Research*, *21*, 193–218.

Wright, D., & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, *62*, 439456.