Cognitively Plausible Models of Human Language Processing

Cognitively Plausible Models of Human Language Processing

Frank Keller
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
keller@inf.ed.ac.uk

Abstract

We pose the development of cognitively plausible models of human language processing as a challenge for computational linguistics. Existing models can only deal with isolated phenomena (e.g., garden paths) on small, specifically selected data sets. The challenge is to build models that integrate multiple aspects of human language processing at the syntactic, semantic, and discourse level. Like human language processing, these models should be incremental, predictive, broad coverage, and robust to noise. This challenge can only be met if standardized data sets and evaluation measures are developed.

1 Introduction

In many respects, human language processing is the ultimate goldstandard for computational linguistics. Humans understand and generate language with amazing speed and accuracy, they are able to deal with ambiguity and noise effortlessly and can adapt to new speakers, domains, and registers. Most surprisingly, they achieve this competency on the basis of limited training data (Hart and Risley, 1995), using learning algorithms that are largely unsupervised.

Given the impressive performance of humans as language processors, it seems natural to turn to psycholinguistics, the discipline that studies human language processing, as a source of information about the design of efficient language processing systems. Indeed, psycholinguists have uncovered an impressive array of relevant facts (reviewed in Section 2), but computational linguists are often not aware of this literature, and results about human language processing rarely inform the design, implementation, or evaluation of artificial language processing systems.

At the same time, research in psycholinguistics is often oblivious of work in computational linguistics (CL). To test their theories, psycholinguists construct computational models of human language processing, but these models often fall short of the engineering standards that are generally accepted in the CL community (e.g., broad coverage, robustness, efficiency): typical psycholinguistic models only deal with isolated phenomena and fail to scale to realistic data sets. A particular issue is evaluation, which is typically anecdotal, performed on a small set of handcrafted examples (see Sections 3).

In this paper, we propose a challenge that requires the combination of research efforts in computational linguistics and psycholinguistics: the development of cognitively plausible models of human language processing. This task can be decomposed into a modeling challenge (building models that instantiate known properties of human language processing) and a data and evaluation challenge (accounting for experimental findings and evaluating against standardized data sets), which we will discuss in turn.

2 Modeling Challenge

2.1 Key Properties

The first part of the challenge is to develop a model that instantiates key properties of human language processing, as established by psycholinguistic experimentation (see Table 1 for an overview and representative references).[1] A striking property of the human language processor is its *efficiency and robustness.* For the vast majority of sentences, it will effortlessly and rapidly deliver the correct analysis, even in the face of noise and ungrammaticalities. There is considerable experimental evi-

[1] Here an in the following, we will focus on sentence processing, which is often regarded as a central aspect of human language processing. A more comprehensive answer to our modeling challenge should also include phonological and morphological processing, semantic inference, discourse processing, and other non-syntactic aspects of language processing. Furthermore, established results regarding the interface between language processing and non-linguistic cognition (e.g., the sensorimotor system) should ultimately be accounted for in a fully comprehensive model.

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers, 60–67.
Uppsala, 2010.

Property	Evidence	Model			
		Rank	Surp	Pred	Stack
Efficiency and robustness	Ferreira et al. (2001); Sanford and Sturt (2002)	−	−	−	+
Broad coverage	Crocker and Brants (2000)	+	+	−	+
Incrementality and connectedness	Tanenhaus et al. (1995); Sturt and Lombardo (2005)	+	+	+	+
Prediction	Kamide et al. (2003); Staub and Clifton (2006)	−	±	+	−
Memory cost	Gibson (1998); Vasishth and Lewis (2006)	−	−	+	+

Table 1: Key properties of human language processing and their instantiation in various models of sentence processing (see Section 2 for details)

dence that shallow processing strategies are used to achieve this. The processor also achieves *broad coverage:* it can deal with a wide variety of syntactic constructions, and is not restricted by the domain, register, or modality of the input.

Human language processing is also word-by-word *incremental*. There is strong evidence that a new word is integrated as soon as it is available into the representation of the sentence thus far. Readers and listeners experience differential processing difficulty during this integration process, depending on the properties of the new word and its relationship to the preceding context. There is evidence that the processor instantiates a strict form of incrementality by building only fully connected trees. Furthermore, the processor is able to make *predictions* about upcoming material on the basis of sentence prefixes. For instance, listeners can predict an upcoming post-verbal element based on the semantics of the preceding verb. Or they can make syntactic predictions, e.g., if they encounter the word *either*, they predict an upcoming *or* and the type of complement that follows it.

Another key property of human language processing is the fact that it operates with limited memory, and that structures in memory are subject to decay and interference. In particular, the processor is known to incur a distance-based *memory cost:* combining the head of a phrase with its syntactic dependents is more difficult the more dependents have to be integrated and the further away they are. This integration process is also subject to interference from similar items that have to be held in memory at the same time.

2.2 Current Models

The challenge is to develop a computational model that captures the key properties of human language processing outlined in the previous section. A number of relevant models have been developed, mostly based on probabilistic parsing techniques, but none of them instantiates all the key properties discussed above (Table 1 gives an overview of model properties).[2]

The earliest approaches were *ranking-based models* (Rank), which make psycholinguistic predictions based on the ranking of the syntactic analyses produced by a probabilistic parser. Jurafsky (1996) assumes that processing difficulty is triggered if the correct analysis falls below a certain probability threshold (i.e., is pruned by the parser). Similarly, Crocker and Brants (2000) assume that processing difficulty ensures if the highest-ranked analysis changes from one word to the next. Both approaches have been shown to successfully model garden path effects. Being based on probabilistic parsing techniques, ranking-based models generally achieve a broad coverage, but their efficiency and robustness has not been evaluated. Also, they are not designed to capture syntactic prediction or memory effects (other than search with a narrow beam in Brants and Crocker 2000).

The ranking-based approach has been generalized by *surprisal models* (Surp), which predict processing difficulty based on the change in the probability distribution over possible analyses from one word to the next (Hale, 2001; Levy, 2008; Demberg and Keller, 2008a; Ferrara Boston et al., 2008; Roark et al., 2009). These models have been successful in accounting for a range of experimental data, and they achieve broad coverage. They also instantiate a limited form of prediction, viz., they build up expectations about the next word in the input. On the other hand, the efficiency and robustness of these models has largely not been evaluated, and memory costs are not modeled (again except for restrictions in beam size).

The *prediction model* (Pred) explicitly predicts syntactic structure for upcoming words (Demberg and Keller, 2008b, 2009), thus accounting for experimental results on predictive language processing. It also implements a strict form of incre-

[2]We will not distinguish between model and linking theory, i.e., the set of assumptions that links model quantities to behavioral data (e.g., more probably structures are easier to process). It is conceivable, for instance, that a stack-based model is combined with a linking theory based on surprisal.

Factor	Evidence
Word senses	Roland and Jurafsky (2002)
Selectional restrictions	Garnsey et al. (1997); Pickering and Traxler (1998)
Thematic roles	McRae et al. (1998); Pickering et al. (2000)
Discourse reference	Altmann and Steedman (1988); Grodner and Gibson (2005)
Discourse coherence	Stewart et al. (2000); Kehler et al. (2008)

Table 2: Semantic factors in human language processing

mentality by building fully connected trees. Memory costs are modeled directly as a distance-based penalty that is incurred when a prediction has to be verified later in the sentence. However, the current implementation of the prediction model is neither robust and efficient nor offers broad coverage.

Recently, a *stack-based model* (Stack) has been proposed that imposes explicit, cognitively motivated memory constraints on the parser, in effect limiting the stack size available to the parser (Schuler et al., 2010). This delivers robustness, efficiency, and broad coverage, but does not model syntactic prediction. Unlike the other models discussed here, no psycholinguistic evaluation has been conducted on the stack-based model, so its cognitive plausibility is preliminary.

2.3 Beyond Parsing

There is strong evidence that human language processing is driven by an interaction of syntactic, semantic, and discourse processes (see Table 2 for an overview and references). Considerable experimental work has focused on the semantic properties of the verb of the sentence, and verb sense, selectional restrictions, and thematic roles have all been shown to interact with syntactic ambiguity resolution. Another large body of research has elucidated the interaction of discourse processing and syntactic processing. The most-well known effect is probably that of referential context: syntactic ambiguities can be resolved if a discourse context is provided that makes one of the syntactic alternatives more plausible. For instance, in a context that provides two possible antecedents for a noun phrase, the processor will prefer attaching a PP or a relative clause such that it disambiguates between the two antecedents; garden paths are reduced or disappear. Other results point to the importance of discourse coherence for sentence processing, an example being implicit causality.

The challenge facing researchers in computational and psycholinguistics therefore includes the development of language processing models that combine syntactic processing with semantic and discourse processing. So far, this challenge is largely unmet: there are some examples of models that integrate semantic processes such as thematic role assignment into a parsing model (Narayanan and Jurafsky, 2002; Padó et al., 2009). However, other semantic factors are not accounted for by these models, and incorporating non-lexical aspects of semantics into models of sentence processing is a challenge for ongoing research. Recently, Dubey (2010) has proposed an approach that combines a probabilistic parser with a model of co-reference and discourse inference based on probabilistic logic. An alternative approach has been taken by Pynte et al. (2008) and Mitchell et al. (2010), who combine a vector-space model of semantics (Landauer and Dumais, 1997) with a syntactic parser and show that this results in predictions of processing difficulty that can be validated against an eye-tracking corpus.

2.4 Acquisition and Crosslinguistics

All models of human language processing discussed so far rely on supervised training data. This raises another aspect of the modeling challenge: the human language processor is the product of an acquisition process that is largely unsupervised and has access to only limited training data: children aged 12–36 months are exposed to between 10 and 35 million words of input (Hart and Risley, 1995). The challenge therefore is to develop a model of language acquisition that works with such small training sets, while also giving rise to a language processor that meets the key criteria in Table 1. The CL community is in a good position to rise to this challenge, given the significant progress in unsupervised parsing in recent years (starting from Klein and Manning 2002). However, none of the existing unsupervised models has been evaluated against psycholinguistic data sets, and they are not designed to meet even basic psycholinguistic criteria such as incrementality.

A related modeling challenge is the development of processing models for languages other than English. There is a growing body of experimental research investigating human language processing in other languages, but virtually all existing psycholinguistic models only work for English (the only exceptions we are aware of are Dubey et al.'s (2008) and Ferrara Boston et al.'s

(2008) parsing models for German). Again, the CL community has made significant progress in crosslinguistic parsing, especially using dependency grammar (Hajič, 2009), and psycholinguistic modeling could benefit from this in order to meet the challenge of developing crosslinguistically valid models of human language processing.

3 Data and Evaluation Challenge

3.1 Test Sets

The second key challenge that needs to be addressed in order to develop cognitively plausible models of human language processing concerns test data and model evaluation. Here, the state of the art in psycholinguistic modeling lags significantly behind standards in the CL community. Most of the models discussed in Section 2 have not been evaluated rigorously. The authors typically describe their performance on a small set of handpicked examples; no attempts are made to test on a range of items from the experimental literature and determine model fit directly against behavioral measures (e.g., reading times). This makes it very hard to obtain a realistic estimate of how well the models achieve their aim of capturing human language processing behavior.

We therefore suggest the development of standard test sets for psycholinguistic modeling, similar to what is commonplace for tasks in computational linguistics: parsers are evaluated against the Penn Treebank, word sense disambiguation systems against the SemEval data sets, co-reference systems against the Tipster or ACE corpora, etc. Two types of test data are required for psycholinguistic modeling. The first type of test data consists of a collection of representative experimental results. This collection should contain the actual experimental materials (sentences or discourse fragments) used in the experiments, together with the behavioral measurements obtained (reading times, eye-movement records, rating judgments, etc.). The experiments included in this test set would be chosen to cover a wide range of experimental phenomena, e.g., garden paths, syntactic complexity, memory effects, semantic and discourse factors. Such a test set will enable the standardized evaluation of psycholinguistic models by comparing the model predictions (rankings, surprisal values, memory costs, etc.) against behavioral measures on a large set of items. This way both the coverage of a model (how many phenomena can it account for) and its accuracy (how well does it fit the behavioral data) can be assessed.

Experimental test sets should be complemented by test sets based on corpus data. In order to assess the efficiency, robustness, and broad coverage of a model, a corpus of unrestricted, naturally occurring text is required. The use of contextualized language data makes it possible to assess not only syntactic models, but also models that capture discourse effects. These corpora need to be annotated with behavioral measures, e.g., eye-tracking or reading time data. Some relevant corpora have already been constructed, see the overview in Table 3, and various authors have used them for model evaluation (Demberg and Keller, 2008a; Pynte et al., 2008; Frank, 2009; Ferrara Boston et al., 2008; Patil et al., 2009; Roark et al., 2009; Mitchell et al., 2010).

However, the usefulness of the psycholinguistic corpora in Table 3 is restricted by the absence of gold-standard linguistic annotation (though the French part of the Dundee corpus, which is syntactically annotated). This makes it difficult to test the accuracy of the linguistic structures computed by a model, and restricts evaluation to behavioral predictions. The challenge is therefore to collect a standardized test set of naturally occurring text or speech enriched not only with behavioral variables, but also with syntactic and semantic annotation. Such a data set could for example be constructed by eye-tracking section 23 of the Penn Treebank (which is also part of Propbank, and thus has both syntactic and thematic role annotation).

In computational linguistics, the development of new data sets is often stimulated by competitions in which systems are compared on a standardized task, using a data set specifically designed for the competition. Examples include the CoNLL shared task, SemEval, or TREC in computational syntax, semantics, and discourse, respectively. A similar competition could be developed for computational psycholinguistics – maybe along the lines of the model comparison challenges that held at the International Conference on Cognitive Modeling. These challenges provide standardized task descriptions and data sets; participants can enter their cognitive models, which were then compared using a pre-defined evaluation metric.[3]

[3]The ICCM 2009 challenge was the Dynamic Stock and Flows Task, for more information see `http://www.hss.cmu.edu/departments/sds/ddmlab/modeldsf/`.

Corpus	Language	Words	Participants	Method	Reference
Dundee Corpus	English, French	50,000	10	Eye-tracking	Kennedy and Pynte (2005)
Potsdam Corpus	German	1,138	222	Eye-tracking	Kliegl et al. (2006)
MIT Corpus	English	3,534	23	Self-paced reading	Bachrach (2008)

Table 3: Test corpora that have been used for psycholinguistic modeling of sentence processing; note that the Potsdam Corpus consists of isolated sentences, rather than of continuous text

3.2 Behavioral and Neural Data

As outlined in the previous section, a number of authors have evaluated psycholinguistic models against eye-tracking or reading time corpora. Part of the data and evaluation challenge is to extend this evaluation to neural data as provided by event-related potential (ERP) or brain imaging studies (e.g., using functional magnetic resonance imaging, fMRI). Neural data sets are considerably more complex than behavioral ones, and modeling them is an important new task that the community is only beginning to address. Some recent work has evaluated models of word semantics against ERP (Murphy et al., 2009) or fMRI data (Mitchell et al., 2008).[4] This is a very promising direction, and the challenge is to extend this approach to the sentence and discourse level (see Bachrach 2008). Again, it will again be necessary to develop standardized test sets of both experimental data and corpus data.

3.3 Evaluation Measures

We also anticipate that the availability of new test data sets will facilitate the development of new evaluation measures that specifically test the validity of psycholinguistic models. Established CL evaluation measures such as Parseval are of limited use, as they can only test the linguistic, but not the behavioral or neural predictions of a model.

So far, many authors have relied on qualitative evaluation: if a model predicts a difference in (for instance) reading time between two types of sentences where such a difference was also found experimentally, then that counts as a successful test. In most cases, no quantitative evaluation is performed, as this would require modeling the reading times for individual item and individual participants. Suitable procedures for performing such tests do not currently exist; linear mixed effects models (Baayen et al., 2008) provide a way of dealing with item and participant variation, but crucially do not enable direct comparisons between models in terms of goodness of fit.

[4]These data sets were released as part of the NAACL-2010 Workshop on Computational Neurolinguistics.

Further issues arise from the fact that we often want to compare model fit for multiple experiments (ideally without reparametrizing the models), and that various mutually dependent measures are used for evaluation, e.g., processing effort at the sentence, word, and character level. An important open challenge is there to develop evaluation measures and associated statistical procedures that can deal with these problems.

4 Conclusions

In this paper, we discussed the modeling and data/evaluation challenges involved in developing cognitively plausible models of human language processing. Developing computational models is of scientific importance in so far as models are implemented theories: models of language processing allow us to test scientific hypothesis about the cognitive processes that underpin language processing. This type of precise, formalized hypothesis testing is only possible if standardized data sets and uniform evaluation procedures are available, as outlined in the present paper. Ultimately, this approach enables qualitative and quantitative comparisons between theories, and thus enhances our understanding of a key aspect of human cognition, language processing.

There is also an applied side to the proposed challenge. Once computational models of human language processing are available, they can be used to predict the difficulty that humans experience when processing text or speech. This is useful for a number applications: for instance, natural language generation would benefit from being able to assess whether machine-generated text or speech is easy to process. For text simplification (e.g., for children or impaired readers), such a model is even more essential. It could also be used to assess the readability of text, which is of interest in educational applications (e.g., essay scoring). In machine translation, evaluating the fluency of system output is crucial, and a model that predicts processing difficulty could be used for this, or to guide the choice between alternative translations, and maybe even to inform human post-editing.

References

ACL. 2010. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala.

Altmann, Gerry T. M. and Mark J. Steedman. 1988. Interaction with context during human sentence processing. *Cognition* 30(3):191–238.

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* to appear.

Bachrach, Asaf. 2008. *Imaging Neural Correlates of Syntactic Complexity in a Naturalistic Context*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Brants, Thorsten and Matthew W. Crocker. 2000. Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken/Luxembourg/Nancy, pages 111–117.

Crocker, Matthew W. and Thorsten Brants. 2000. Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research* 29(6):647–669.

Demberg, Vera and Frank Keller. 2008a. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 101(2):193–210.

Demberg, Vera and Frank Keller. 2008b. A psycholinguistically motivated version of TAG. In *Proceedings of the 9th International Workshop on Tree Adjoining Grammars and Related Formalisms*. Tübingen, pages 25–32.

Demberg, Vera and Frank Keller. 2009. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In Taatgen and van Rijn (2009), pages 1888–1893.

Dubey, Amit. 2010. The influence of discourse on syntax: A psycholinguistic model of sentence processing. In ACL.

Dubey, Amit, Frank Keller, and Patrick Sturt. 2008. A probabilistic corpus-based model of syntactic parallelism. *Cognition* 109(3):326–344.

EMNLP. 2009. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Singapore.

Ferrara Boston, Marisa, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research* 2(1):1–12.

Ferreira, Fernanda, Kiel Christianson, and Andrew Hollingworth. 2001. Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research* 30(1):3–20.

Frank, Stefan L. 2009. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In Taatgen and van Rijn (2009), pages 1139–1144.

Garnsey, Susan M., Neal J. Pearlmutter, Elisabeth M. Myers, and Melanie A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language* 37(1):58–93.

Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68:1–76.

Grodner, Dan and Edward Gibson. 2005. Consequences of the serial nature of linguistic input. *Cognitive Science* 29:261–291.

Hajič, Jan, editor. 2009. *Proceedings of the 13th Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, Boulder, CO.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Pittsburgh, PA, volume 2, pages 159–166.

Hart, Betty and Todd R. Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Paul H. Brookes, Baltimore, MD.

Jurafsky, Daniel. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2):137–194.

Kamide, Yuki, Gerry T. M. Altmann, and Sarah L. Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language* 49:133–156.

Kehler, Andrew, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman. 2008. Coherence and coreference revisited. *Journal of Semantics* 25(1):1–44.

Kennedy, Alan and Joel Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision Research* 45:153–168.

Klein, Dan and Christopher Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, pages 128–135.

Kliegl, Reinhold, Antje Nuthmann, and Ralf Engbert. 2006. Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General* 135(1):12–35.

Landauer, Thomas K. and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2):211–240.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3):1126–1177.

McRae, Ken, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language* 38(3):283–312.

Mitchell, Jeff, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In ACL.

Mitchell, Tom M., Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just3. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320(5880):1191–1195.

Murphy, Brian, Marco Baroni, and Massimo Poesio. 2009. EEG responds to conceptual stimuli and corpus semantics. In EMNLP, pages 619–627.

Narayanan, Srini and Daniel Jurafsky. 2002. A Bayesian model predicts human parse preference and reading time in sentence processing. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA, pages 59–65.

Padó, Ulrike, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science* 33(5):794–838.

Patil, Umesh, Shravan Vasishth, and Reinhold Kliegl. 2009. Compound effect of probabilistic disambiguation and memory retrievals on sentence processing: Evidence from an eye-tracking corpus. In A. Howes, D. Peebles, and R. Cooper, editors, *Proceedings of 9th International Conference on Cognitive Modeling*. Manchester.

Pickering, Martin J. and Martin J. Traxler. 1998. Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning Memory and Cognition* 24(4):940–961.

Pickering, Martin J., Matthew J. Traxler, and Matthew W. Crocker. 2000. Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language* 43(3):447–475.

Pynte, Joel, Boris New, and Alan Kennedy. 2008. On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research* 48(21):2172–2183.

Roark, Brian, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In EMNLP, pages 324–333.

Roland, Douglas and Daniel Jurafsky. 2002. Verb sense and verb subcategorization probabilities. In Paola Merlo and Suzanne Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, John Bejamins, Amsterdam, pages 325–346.

Sanford, Anthony J. and Patrick Sturt. 2002. Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences* 6:382–386.

Schuler, William, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-coverage parsing using human-like memory constraints. *Computational Linguistics* 26(1):1–30.

Staub, Adrian and Charles Clifton. 2006. Syntactic prediction in language comprehension: Evidence from either . . . or. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32:425–436.

Stewart, Andrew J., Martin J. Pickering, and Anthony J. Sanford. 2000. The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language* 42(3):423–443.

Sturt, Patrick and Vincenzo Lombardo. 2005. Processing coordinated structures: Incrementality and connectedness. *Cognitive Science* 29(2):291–305.

Taatgen, Niels and Hedderik van Rijn, editors. 2009. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Amsterdam.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268:1632–1634.

Vasishth, Shravan and Richard L. Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language* 82(4):767–794.