A Model of Discourse Predictions in Human Sentence Processing

Amit Dubey and **Frank Keller** and **Patrick Sturt**
Human Communication Research Centre, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
{amit.dubey,frank.keller,patrick.sturt}@ed.ac.uk

Abstract

This paper introduces a psycholinguistic model of sentence processing which combines a Hidden Markov Model noun phrase chunker with a co-reference classifier. Both models are fully incremental and generative, giving probabilities of lexical elements conditional upon linguistic structure. This allows us to compute the information theoretic measure of surprisal, which is known to correlate with human processing effort. We evaluate our surprisal predictions on the Dundee corpus of eye-movement data show that our model achieve a better fit with human reading times than a syntax-only model which does not have access to co-reference information.

1 Introduction

Recent research in psycholinguistics has seen a growing interest in the role of *prediction* in sentence processing. Prediction refers to the fact that the human sentence processor is able to anticipate upcoming material, and that processing is facilitated when predictions turn out to be correct (evidenced, e.g., by shorter reading times on the predicted word or phrase). Prediction is presumably one of the factors that contribute to the efficiency of human language understanding. Sentence processing is *incremental* (i.e., it proceeds on a word-by-word basis); therefore, it is beneficial if unseen input can be anticipated and relevant syntactic and semantic structure constructed in advance. This allows the processor to save time and makes it easier to cope with the constant stream of new input.

Evidence for prediction has been found in a range of psycholinguistic processing domains. *Semantic prediction* has been demonstrated by studies that show anticipation based on selectional restrictions: listeners are able to launch eye-movements to the predicted argument of a verb before having encountered it, e.g., they will fixate an edible object as soon as they hear the word *eat* (Altmann and Kamide, 1999). Semantic prediction has also been shown in the context of semantic priming: a word that is preceded by a semantically related prime or by a semantically congruous sentence fragment is processed faster (Stanovich and West, 1981; Clifton et al., 2007). An example for *syntactic prediction* can be found in coordinate structures: readers predict that the second conjunct in a coordination will have the same syntactic structure as the first conjunct (Frazier et al., 2000). In a similar vein, having encountered the word *either*, readers predict that *or* and a conjunct will follow it (Staub and Clifton, 2006). Again, priming studies corroborate this: Comprehenders are faster at naming words that are syntactically compatible with prior context, even when they bear no semantic relationship to it (Wright and Garrett, 1984).

Predictive processing is not confined to the sentence level. Recent experimental results also provide evidence for *discourse prediction*. An example is the study by van Berkum et al. (2005), who used a context that made a target noun highly predictable, and found a mismatch effect in the ERP (event-related brain potential) when an adjective appeared that was inconsistent with the target noun. An example is (we give translations of their Dutch materials):

(1) The burglar had no trouble locating the secret family safe.

a. Of course, it was situated behind a

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 304–312,
Edinburgh, Scotland, UK, July 27–31, 2011.

big$_{neu}$ but unobtrusive painting$_{neu}$.

b. Of course, it was situated behind a big$_{com}$ but unobtrusive bookcase$_{com}$.

Here, the adjective *big*, which can have neutral or common gender in Dutch, is consistent with the predicted noun *painting* in (1-a), but inconsistent with it in (1-b), leading to a mismatch ERP on *big* in (1-b) but not in (1-a).

Previous results on discourse effects in sentence processing can also be interpreted in terms of prediction. In a classical paper, Altmann and Steedman (1988) demonstrated that PP-attachment preferences can change through discourse context: if the context contains two potential referents for the target NP, then NP-attachment of a subsequent PP is preferred (to disambiguate between the two referents), while if the context only contains one target NP, VP-attachment is preferred (as there is no need to disambiguate). This result (and a large body of related findings) is compatible with an interpretation in which the processor predicts upcoming syntactic attachment based on the presence of referents in the preceding discourse.

Most attempts to model prediction in human language processing have focused on syntactic prediction. Examples include Hale's (2001) surprisal model, which relates processing effort to the conditional probability of the current word given the previous words in the sentence. This approach has been elaborated by Demberg and Keller (2009) in a model that explicitly constructs predicted structure, and includes a verification process that incurs additional processing cost if predictions are not met. Recent work has attempted to integrate semantic and discourse prediction with models of syntactic processing. This includes Mitchell et al.'s (2010) approach, which combines an incremental parser with a vector-space model of semantics. However, this approach only provides a loose integration of the two components (through simple addition of their probabilities), and the notion of semantics used is restricted to lexical meaning approximated by word co-occurrences. At the discourse level, Dubey (2010) has proposed a model that combines an incremental parser with a probabilistic logic-based model of co-reference resolution. However, this model does not explicitly model discourse effects in terms of prediction, and again only proposes a loose integration of co-reference and syntax. Furthermore, Dubey's (2010) model has only been tested on two experimental data sets (pertaining to the interaction of ambiguity resolution with context), no broad coverage evaluation is available.

The aim of the present paper is to overcome these limitations. We propose a computational model that captures discourse effects on syntax in terms of prediction. The model comprises a co-reference component which explicitly stores discourse mentions of NPs, and a syntactic component which adjust the probabilities of NPs in the syntactic structure based on the mentions tracked by the discourse component. Our model is HMM-based, which makes it possible to efficiently process large amounts of data, allowing an evaluation on eye-tracking corpora, which has recently become the gold-standard in computational psycholinguistics (e.g., Demberg and Keller 2008; Frank 2009; Boston et al. 2008; Mitchell et al. 2010).

The paper is structured as follows: In Section 2, we describe the co-reference and the syntactic models and evaluate their performance on standard data sets. Section 3 presents an evaluation of the overall model on the Dundee eye-tracking corpus. The paper closes with a comparison with related work and a general discussion in Sections 4 and 5.

2 Model

This model utilises an NP chunker based upon a hidden Markov model (HMM) as an approximation to syntax. Using a simple model such as an HMM facilitates the integration of a co-reference component, and the fact that the model is generative is a prerequisite to using surprisal as our metric of interest (as surprisal require the computation of prefix probabilities). The key insight in our model is that human sentence processing is, on average, facilitated when a previously-mentioned discourse entity is repeated. This facilitation depends upon keeping track of a list of previously-mentioned entities, which requires (at the least) shallow syntactic information, yet the facilitation itself is modeled primarily as a lexical phenomenon. This allows a straightforward separation of concerns: shallow syntax is captured using the HMM's hidden states, whereas the co-reference facilitation is modeled using the HMM's emissions. The vocabulary of hidden states is described in Section 2.1 and the emission distribution in Section 2.2

2.1 Syntactic Model

A key feature of the co-reference component of our model (described below) is that syntactic analysis and co-reference resolution happen simultaneously. This could potentially slow down the syntactic analysis, which tends to already be quite slow for exhaustive surprisal-based incremental parsers. Therefore, rather than using full parsing, we use an HMM-based NP chunker which allows for a fast analysis. NP chunking is sufficient to extract NP discourse mentions and, as we show below, surprisal values computed using HMM chunks provide a useful fit on the Dundee eye-movement data.

To allow the HMM to handle possessive constructions as well as NP with simple modifiers and complements, the HMM decodes NP subtrees with depth of 2, by encoding the start, middle and end of a syntactic category X as '(X', 'X' and 'X)', respectively. To reduce an explosion in the number of states, the category begin state '(X' only appears at the rightmost lexical token of the constituent's leftmost daughter. Likewise, 'X)' only appears at the leftmost lexical token of the constituent's rightmost daughter. An example use of this state vocabulary can be seen in Figure 1. Here, a small degree of recursion allows for the NP ((new york city's) general obligation fund) to be encoded, with the outer NP's left bracket being 'announced' at the token *'s*, which is the rightmost lexical token of the inner NP. Hidden states also include part-of-speech (POS) tags, allowing simultaneous POS tagging. In the example given in Figure 1, the full state can be read by listing the labels written above a word, from top to bottom. For example, the full state associated with *'s* is (NP-NP)-POS. As *'s* can also be a contraction of *is*, another possible state for *'s* is VBZ (without recursive categories as we are only interested in NP chunks).

The model uses unsmoothed bi-gram transition probabilities, along with a maximum entropy distribution to guess unknown word features. The resulting distribution has the form $P(tag|word)$ and is therefore unsuitable for computing surprisal values. However, using Bayes' theorem we can compute:

$$P(word|tag) = \frac{P(tag|word)P(word)}{P(tag)} \qquad (1)$$

which is what we need for surprisal. The primary information from this probability comes from $P(tag|word)$, however, reasonable estimates of $P(tag)$ and $P(word)$ are required to ensure the probability distribution is proper. $P(tag)$ may be estimated on a parsed treebank. $P(word)$, the probability of a particular unseen word, is difficult to estimate directly. Given that our training data contains approximately 10^6 words, we assume that this probability must be bounded above by 10^{-6}. As an approximation, we use this upper bound as the probability of $P(word)$.

Training The chunker is trained on sections 2–22 of the Wall Street Journal section of the Penn Treebank. CoNLL 2000 included chunking as a shared task, and the results are summarized by Tjong Kim Sang and Buchholz (2000). Our chunker is not comparable to the systems in the shared task for several reasons: we use more training data, we tag simultaneously (the CoNLL systems used gold standard tags) and our notion of a chunk is somewhat more complex than that used in CoNLL. The best performing chunker from CoNLL 2000 achieved an F-score of 93.5%, and the worst performing system an F-score of 85.8%. Our chunker achieves a comparable F-score of 85.5%, despite the fact that it simultaneously tags and chunks, and only uses a bigram model.

2.2 Co-Reference Model

In a standard HMM, the emission probabilities are computed as $P(w_i|s_i)$ where w_i is the i^{th} word and s_i is the i^{th} state. In our model, we replace this with a choice between two alternatives:

$$P(w_i|s_i) = \begin{cases} \lambda P_{\text{seen before}}(w_i|s_i) \\ (1-\lambda)P_{\text{discourse new}}(w_i|s_i) \end{cases} \qquad (2)$$

The 'discourse new' probability distribution is the standard HMM emission distribution. The 'seen before' distribution is more complicated. It is in part based upon caching language models. However, the contents of the cache are not individual words but

						(NP	NP	NP	NP)					
(NP	NP)		(NP	NP	NP	NP)	NP					(NP	NP	NP)
JJ	NN	IN	NNP	NNP	NNP	POS	JJ	NN	NNS	VBN	RP	DT	NN	NN
strong	demand	for	new	york	city	’s	general	obligation	bonds	propped	up	the	municipal	market

Figure 1: The chunk notation of a tree from the training data.

Variable	Type
l, l'	List of trie nodes
w, w_i	Words
t	Tag
n, n'	Trie nodes

```
$l \leftarrow List$(root of mention trie)
for $w \leftarrow w_0$ to $w_n$ do
    $l' \leftarrow l$
    $l \leftarrow \emptyset$
    Clear tag freq array $ft$
    Clear word freq array $fwt$
    for $t \in$ tag set do
        for $n \in l'$ do
            $ft(t) \leftarrow ft(t) + FreqOf(n,t)$
            $n' \leftarrow Getchild(w,t)$
            if $n' \neq \emptyset$ then
                $fwt(t) \leftarrow fwt(t) + FreqOf(n',w,t)$
                $l \leftarrow n' :: l$
            end if
        end for
    end for
    $P_{\text{seen before}}(w|t) = ft(t)/fwt(t)$
end for
```

Figure 2: Looking up entries from the NP Cache

rather a collection of all NPs mentioned so far in the document.

Using a collection of NPs rather than individual words complicates the decoding process. If m is the size of a document, and n is the size of the current sentence, decoding occurs in $O(mn)$ time as opposed to $O(n)$, as the collection of NPs needs to be accessed at each word. However, we do not store the NPs in a list, but rather a trie. This allows decoding to occur in $O(n \log m)$ time, which we have found to be quite fast in practise. The algorithm used to keep track of currently active NPs is presented in Figure 2. This shows how the distribution $P_{\text{seen before}}$ is updated on a word-by-word basis. At the end of each sentence, the NPs of the Viterbi parse are added to the mention trie after having their leading articles stripped. A weakness of the algorithm is that mentions are only added on a sentence-by-sentence basis (disallowing within-sentence references). Although the algorithm is intended to find whole-string matches, in practise, it will count any NP whose prefix matches as being co-referent.

A consequence of Equation 2 is that co-reference resolution is handled at the same time as HMM decoding. Whenever the ‘seen before’ distribution is applied, an NP is co-referent with one occurring earlier. Likewise, whenever the ‘discourse new’ distribution is applied, the NP is not co-referent with any NP appearing previously. As one choice or the other is made during decoding, the decoder therefore also selects a chain of co-referent entities. Generally, for words which *have* been used in this discourse, the magnitude of probabilities in the ‘seen before’ distribution are much higher than in the ‘discourse new’ distribution. Thus, there is a strong bias to classify NPs which match word-for-word as being co-referent. There remains a possibility that the model primarily captures lexical priming, rather than co-reference. However, we note that string match is a strong indicator of two NPs being coreferent (cf. Soon et al. 2001), and, moreover, the matching is done on an NP-by-NP basis, which is more suitable for finding entity coreference, rather than a word-by-word basis, which would be more suitable for lexical priming.

An appealing side-effect of using a simple co-reference decision rule which is applied incrementally is that it is relatively simple to incrementally compute the transitive closure of co-reference chains, resulting in the entity sets which are then used in evaluation.

The co-reference model only has one free parameter, λ, which is estimated from the ACE-2 corpus. The estimate is computed by counting how often a repeated NP actually is discourse new. In the current implementation of the model, λ is constant throughout the test runs. However, λ could possibly be a function of the previous discourse, allowing for more complicated classification probabilities.

3 Evaluation

3.1 Data

Our evaluation experiments were conducted upon the Dundee corpus (Kennedy et al., 2003), which contains the eye-movement record of 10 participants each reading 2,368 sentences of newspaper text. This data set has previously been used by Demberg and Keller (2008) and Frank (2009) among others.

3.2 Evaluation

Eye tracking data is noisy for a number of reasons, including the fact that experimental participants can look at any word which is currently displayed. While English is normally read in a left-to-right manner, readers often skip words or make regressions (i.e., look at a word to the left of the one they are currently fixating). Deviations from a strict left-to-right progression of fixations motivate the need for several different measures of eye movement. The model presented here predicts the Total Time that participants spent looking at a region, which includes any re-fixations after looking away. In addition to total time, other possible measures include (a) First Pass, which measures the initial fixation and any re-fixations before looking at any other word (this occurs, for instance, if the eye initially lands at the start of a long word – the eye will tend to re-fixate on a more central viewing location), (b) Right Bounded reading time, which includes all fixations on a word before moving to the right of the word (i.e., re-fixations after moving left are included), and (c) Second Pass, which includes any re-fixation on a word after looking at any other word (be it to the left or the right of the word of interest). We found that the model performed similarly across all these reading time metrics, we therefore only report results for Total Time.

As mentioned above, reading measures are hypothesised to correlate with Surprisal, which is defined as:

$$S(w_t) = -\log(P(w_t|w_1...w_{t1})) \tag{3}$$

We compute the surprisal scores for the syntax-only HMM, which does not have access to co-reference information (henceforth referred to as 'HMM') and the full model, which combines the syntax-only HMM with the co-reference model (henceforth 'HMM+Ref'). To determine if our Dundee corpus simulations provide a reasonable model of human sentence processing, we perform a regression analysis with the Dundee corpus reading time measure as the dependent variable and the surprisal scores as the independent variable.

To account for noise in the corpus, we also use a number of additional explanatory variables which are known to strongly influence reading times. These include the logarithm of the frequency of a word (measured in occurrences per million) and the length of a word in letters. Two additional explanatory variables were available in the Dundee corpus, which we also included in the regression model. These were the position of a word on a line, and which line in a document a word appeared in. As participants could only view one line at a time (i.e., one line per screen), these covariates are known as line position and screen position, respectively.

All the covariates, including the surprisal estimates, were centered before including them in the regression model. Because the HMM and HMM+Ref surprisal values are highly collinear, the HMM+Ref surprisal values were added as residuals of the HMM surprisal values.

In a normal regression analysis, one must either assume that participants or the particular choice of items add some randomness to the experiment, and either each participant's responses for all items must be averaged (treating participants as a random factor), or all participant's responses for each item is averaged (treating items as a random factor). However, in the present analysis we utilise a mixed effects model, which allows both items and participants to be treated as random factors.[1]

The are a number of criteria which can be used to test the efficacy of one regression model over another. These include the Aikake Information Criterion (AIC), the Bayesian Information Criterion (BIC), which trade off model fit and number of model parameters (lower scores are better). It is also common to compare the log-likelihood of the models (higher log-likelihood is better), in which case a χ^2 can be used to evaluate if a model offers a significantly better fit, given the number of parameters is uses. We test three models: (i) a baseline, with only low-level factors as independent variables; (ii) the HMM model, with the baseline factors plus surprisal computed by the syntax-only HMM; and (iii) the HMM+Ref model which includes the raw surprisal values of the syntax-only HMM and the surprisal of the HMM+Ref models as computed as a residual of the HMM surprisal score. We compare the HMM and HMM+Ref to the baseline, and the HMM+Ref model against the HMM model.

Some of the data needed to be trimmed. If, due to data sparsity, the surprisal of a word goes to infinity for one of the models, we entirely remove that word from the analysis. This occurred seven times form the HMM+Ref model, but did not occur at all with the HMM model. Some of the eye-movement data was trimmed, as well. Fixations on the first and last words of a line were excluded, as were tracklosses. However, we did not trim any items due to abnormally short or abnormally long fixation durations.

3.3 Results

The result of the model comparison on Total Time reading data is summarised in Table 1. To allow this work to be compared with other models, the lower part of the table gives the abosolute AIC, BIC and log likelihood of the baseline model, while the upper part gives delta AIC, BIC and log likelihood scores of pairs of models.

We found that both the HMM and HMM+Ref provide a significantly better fit with the reading time data than the Baseline model; all three criteria agree: AIC and BIC lower than for the baseline, and log-likelihood is higher. Moreover, the HMM+Ref model provides a significantly better fit than the HMM model, which demonstrates the benefit of co-reference information for modeling reading times. Again, all three measures provide the same result.

Table 2 corroborates this result. It list the mixed-model coefficients for the HMM+Ref model and shows that all factors are significant predictors, including both HMM surprisal and residualized HMM+Ref surprisal.

4 Related Work

There have been few computational models of human sentence processing that have incorporated a referential or discourse-level component. Niv (1994) proposed a parsing model based on Combinatory Categorial Grammar (Steedman, 2001), in which referential information was used to resolve syntactic ambiguities. The model was able to capture effects of referential information on syntactic garden paths (Altmann and Steedman, 1988). This model differs from that proposed in the present paper, as it is intended to capture psycholinguistic preferences in a qualitative manner, whereas the aim of the present model is to provide a *quantitative* fit to measures of processing difficulty. Moreover, the model was not based on a large-scale grammar, and was not tested on unrestricted text. Spivey and Tanenhaus (1998) proposed a sentence processing model that examined the effects of referential information, as well as other constraints, on the resolution of ambiguous sentences. Unlike Niv (1994),

[1] We assume that each participant and item bias the reading time of the experiment. Such an analysis is known as having random intercepts of participant and item. It is also possible to assume a more involved analysis, known as random slopes, where the participants and items bias the slope of the predictor. The model did not converge when using random intercept and slopes on both participant and item. If random slopes on items were left out, the HMM regression model did converge, but not the HMM+Ref model. As the HMM+Ref is the model of interest random slopes were left out entirely to allow a like-with-like comparison between the HMM and HMM+Ref regression models.

From	To	Δ AIC	Δ BIC	Δ logLik	χ^2	Significance
Baseline	HMM	-80	-69	41	82.112	$p < .001$
Baseline	HMM+Ref	-99	-89	51	101.54	$p < .001$
HMM	HMM+Ref	-19	-8	11	21.424	$p < .001$

Model	AIC	BIC	logLik
Baseline	10567789	10567880	-5283886

Table 1: Model comparison (upper part) and absolute scores for the Baseline model (lower part)

Coefficient	Estimate	Std Error	t-value
(Intercept)	991.4346	23.7968	41.66
log(Word Frequency)	-55.3045	1.4830	-37.29
Word Length	128.6216	1.4677	87.63
Screen Position	-1.7769	0.1326	-13.40
Line Position	10.1592	0.7387	13.75
HMM	12.1287	1.3366	9.07
HMM+Ref	19.2772	4.1627	4.63

Table 2: Coefficients of the HMM+Ref model on Total Reading Times. Note that $t > 2$ indicates that the factor in question is a significant predictor.

Spivey and Tanenhaus's (1998) model was specifically designed to provide a quantitative fit to reading times. However, the model lacked generality, being designed to deal with only one type of sentence. In contrast to both of these earlier models, the model proposed here aims to be general enough to provide estimated reading times for unrestricted text. In fact, as far as we are aware, the present paper represents the first wide-coverage model of human parsing that has incorporated discourse-level information.

5 Discussion

The primary finding of this work is that incorporating discourse information such as co-reference into an incremental probabilistic model of sentence processing has a beneficial effect on the ability of the model to predict broad-coverage human parsing behaviour.

Although not thoroughly explored in this paper, our finding is related to an ongoing debate about the structure of the human sentence processor. In particular, the model of Dubey (2010), which also simulates the effect of discourse on syntax, is aimed at examining *interactivity* in the human sentence processor. Interactivity describes the degree to which human parsing is influenced by non-syntactic factors. Under the *weakly interactive* hypothesis, discourse factors may prune or re-weight parses, but only when assuming the *strongly interactive* hypothesis would we argue that the sentence processor predicts upcoming material due to discourse factors. Dubey found that a weakly interactive model simulated a pattern of results in an experiment (Grodner et al., 2005) which was previously believed to provide evidence for the strongly interactive hypothesis. However, as Dubey does not provide broad-coverage parsing results, this leaves open the possibility that the model cannot generalise beyond the experiments expressly modeled in Dubey (2010).

The model presented here, on the other hand, is not only broad-coverage but could also be described as a strongly interactive model. The strong interactivity arises because co-reference resolution is strongly tied to lexical generation probabilities, which are part of the syntactic portion of our model. This cannot be achieve in a weakly interactive model, which is limited to pruning or re-weighting of parses based on discourse information. As our analysis on the Dundee corpus showed, the lexical probabilities (in the form of HMM+Ref surprisal) are key to improving the fit on eye-tracking data. We therefore argue that our results provide evidence against a weakly interactive approach, which may be sufficient to model individual phenomena (as shown by Dubey 2010), but is unlikely to be able to match the broad-coverage result we have presented here. We also note that psycholinguistic evidence for discourse prediction (such as the context based lexical prediction shown by van Berkum et al. 2005, see Section 1) is also evidence for strong interactivity; prediction goes beyond mere pruning or reweighting and requires strong interactivity.

References

Gerry Altmann and Mark Steedman. Interaction with context during human sentence processing. *Cognition*, 30:191–238, 1988.

Gerry T. M. Altmann and Yuki Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–264, 1999.

Marisa Ferrara Boston, John T. Hale, Reinhold Kliegl, and Shravan Vasisht. Surprising parser actions and reading difficulty. In *Proceedings of ACL-08:HLT, Short Papers*, pages 5–8, 2008.

Charles Clifton, Adrian Staub, and Keith Rayner. Eye movement in reading words and sentences. In R V Gompel, M Fisher, W Murray, and R L Hill, editors, *Eye Movements: A Window in Mind and Brain*, pages 341–372. Elsevier, 2007.

Vera Demberg and Frank Keller. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109: 192–210, 2008.

Vera Demberg and Frank Keller. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the 29th meeting of the Cognitive Science Society (CogSci-09)*, 2009.

Amit Dubey. The influence of discourse on syntax: A psycholinguistic model of sentence processing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, 2010.

Stefan Frank. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *31st Annual Conference of the Cognitive Science Society (COGSCI 2009)*, Amsterdam, The Netherlands, 2009.

Lyn Frazier, Alan Munn, and Charles Clifton. Processing coordinate structure. *Journal of Psycholinguistic Research*, 29:343–368, 2000.

Daniel J. Grodner, Edward A. F. Gibson, and Duane Watson. The influence of contextual constrast on syntactic processing: Evidence for strong-interaction in sentence comprehension. *Cognition*, 95(3):275–296, 2005.

John T. Hale. A probabilistic earley parser as a psycholinguistic model. In *In Proceedings of the Second Meeting of the North American Chapter of the Asssociation for Computational Linguistics*, 2001.

A. Kennedy, R. Hill, and J. Pynte. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010.

M. Niv. A psycholinguistically motivated parser for CCG. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 125–132, Las Cruces, NM, 1994.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27 (4):521–544, 2001.

M. J. Spivey and M. K. Tanenhaus. Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24(6):1521–1543, 1998.

Kieth E. Stanovich and Richard F. West. The effect of sentence context on ongoing word recognition: Tests of a two-pricess theory. *Journal of Experimental Psychology: Human Perception and Performance*, 7:658–672, 1981.

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

Mark Steedman. *The Syntactic Process*. Bradford Books, 2001.

Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of CoNLL-2000 and LLL-2000*, pages 127–132. Lisbon, Portugal, 2000.

Jos J. A. van Berkum, Colin M. Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31(3):443–467, 2005.

Barton Wright and Merrill F. Garrett. Lexical decision in sentences: Effects of syntactic structure. *Memory and Cognition*, 12:31–45, 1984.