Précis of Thesis: A Broad-Coverage Model of Prediction in Human Sentence Processing

Vera Demberg-Winterfors

The University of Edinburgh
supervised by Dr. Frank Keller and Prof. Fernanda Ferreira

1 Introduction

The aim of this thesis is to design and implement a cognitively plausible theory of sentence processing which incorporates a mechanism for modelling a prediction and verification process in human language understanding, and to evaluate the validity of this model on specific psycholinguistic phenomena as well as on broad-coverage, naturally occurring text. "Prediction" in this context means that words or categories are anticipated based on previously processed words.

Modeling prediction is a timely and relevant contribution to the field of psycholinguistics because recent experimental evidence suggests that humans predict upcoming syntactic structure or words during sentence processing. However, none of the current sentence processing theories capture prediction explicitly. This thesis proposes a novel model of word-by-word incremental sentence processing that offers an explicit prediction and verification mechanism.

In evaluating the proposed model on broad-coverage naturalistic text, this thesis also makes a methodological contribution. The design and evaluation of current sentence processing theories are usually based exclusively on experimental results from individual psycholinguistic experiments on specific linguistic structures. However, a theory of language processing in humans should not only work in an experimentally designed environment, but should also have explanatory power for naturally occurring language.

Interdisciplinary Contribution

This thesis is strongly interdisciplinary in nature; its methods are drawn from, and it makes contributions to, three different fields: cognitive psychology (psycholinguistics), artificial intelligence (natural language processing) and linguistics (formal grammar).

The contribution to psycholinguistics is twofold. It consists of pioneering the evaluation of sentence processing theories on a broad-coverage, naturalistic corpus (this contribution has been recognized by the field through the "AMLaP Young Scientist Award").

Additionally, this thesis fills an important gap in the development of psycholinguistic sentence processing theories, as it is the first theory of syntax processing which explicitly models the prediction (in terms of anticipating upcoming words and structure) which has recently been observed in human sentence comprehension (e.g., Kamide et al., 2003; van Berkum et al., 2005; Staub and Clifton, 2006). Our theory furthermore accounts for more of the sentence processing phenomena than other current sentence processing theories and best predicts the variation observed in reading times on naturalistic broad-coverage text. It is hence not only supported by psycholinguistically plausible underlying mechanisms but also by strong empirical evidence, and is thus a step forward in gaining a better understanding of the mechanisms involved in human cognition.

The sentence processing theory developed in this thesis also constitutes a contribution to the area of natural language processing: A theory that scales to broad-coverage text processing and can adequately assess where human processing difficulty arises is also of high interest in many computational linguistics applications, in particular for applications that generate text or speech, and need to optimize the understandability of the generated linguistic output, such as in dialogue systems, readability assessments, teaching and tutoring applications, text summarization and text simplification etc.

A further contribution to natural language processing consists of the strictly incremental parser developed as part of this thesis. The parser can be integrated with time-critical language processing applications, where processing is critical to proceed incrementally as the sentence unfolds. Only a strictly incremental parser spells out all relationships between all perceived words and hence allows for the largest degree of incremental interpretation, which in turn allows for example for faster speech-to-speech translation and more immediate reactions to instructions (for example by a speech-driven agent or robot).

The thesis contributes to the study of formal grammar in that it develops a novel, psycholinguistically motivated version of tree-adjoining grammar, which supports strict incrementality and prediction.

In the following, we will give an overview of the main contributions of this work (Chapters 3-9 of the thesis).

2 Proof of Concept for Evaluation on Naturalistic Data

The first goal in this thesis relates to the evaluation of sentence processing theories on naturalistic text. We need to show that a corpus of naturalistic text constitutes a valid and valuable resource for testing sentence processing theories. Our resource of naturalistic text is the Dundee Corpus, a collection of 20 newspaper articles comprising roughly 50,000 words, which was annotated with the eye-movements of 10 readers. The Dundee Corpus is analysed in detail in Chapter 3, in particular also including a discussion of particularities of running mixed-effects regression models on such naturalistic data as opposed to working with data from experimental materials. Chapter 4 investigates whether a benchmark processing effect, the subject relative clause (SRC) vs. object relative clause (ORC) asymmetry, can be detected in this data set. The SRC/ORC asymmetry effect refers to the finding that English subject relative clauses (SRCs) as in (1-a) are easier to process than object relative clauses (ORCs) as in (1-b). Experimentally, this difficulty is evidenced by the fact that reading times on region R1 in the SRC are lower than reading times for the corresponding region R2 in the ORC (e.g., King and Just, 1991).

(1) a. The reporter who $[\text{attacked}]_{R1}$ the senator admitted the error.
b. The reporter who the senator $[\text{attacked}]_{R2}$ admitted the error.

The ORC difficulty effect is explained by processing theories that capture the complexity involved in computing the syntactic dependencies between the words in a sentence. The most prominent such theory is Dependency Locality Theory (DLT), proposed by Gibson (1998). We automatically extracted (and manually checked for correctness) all relative clauses from the Dundee corpus, and computed mixed-effects linear regression models to determine whether reading times were higher on the embedded verbs of object relative clauses than on the embedded verbs of subject relative clauses. Our regression results show that the difference between subject and object relative clauses, measured in terms of DLT integration cost at the embedded verb, is a significant positive predictor of reading times.

The fact that such a well-known laboratory effect can be replicated on the naturally occurring text suggests that the validity of sentence processing theories, which was previously only tested on data obtained for isolated, manually constructed sentences in controlled lab experiments, can be enhanced considerably if we are able to show that they scale up to model reading data from an eye-tracking corpus of naturally occurring text.

3 Evaluating two previous Sentence Processing Theories on Broad-Coverage, Naturalistic Data

Chapter 5 evaluates two existing well-established theories of sentence processing, Surprisal (Hale, 2001; Levy, 2008) and Dependency Locality Theory (DLT; Gibson, 1998)[1], on the full Dundee corpus. Surprisal and DLT were chosen among a range of alternative sentence processing theories, because they are prominent in the field, are supported by a good range of empirical data and in addition make complimentary assumptions about the source of processing difficulty: DLT's integration cost captures the cost incurred when a head has to be integrated with the dependents that precede it, with more difficulty being encountered if a larger number of discourse referents has occurred in between the dependent and its head. Surprisal, on the other hand, accounts for the cost that results when the current word is not compatible with the most likely analyses of the preceding context, i.e. when it is unexpected, which can also be thought of as a word being more difficult if it carries a lot of information. Integration cost can hence be regarded as a backward looking cost (past material has to be held in memory and integrated), while Surprisal can be thought of as a forward-looking cost (unexpected events cause processing difficulty because any syntactic analyses not compatible with the current word have to be discarded).

Processing difficulty estimates for both theories were calculated automatically for each word in the corpus. The processing difficulty calculations are based on the Roark et al. (2009) parser to determine Surprisal estimates, and on the MINIPAR parser (Lin, 1998) for DLT integration costs. We then used linear mixed-effects regression models to determine whether the difficulty predictions can account for any of the variance in the reading time data (which is not already accounted for by other more basic parameters known to influence reading times). This evaluation constitutes the first broad-coverage comparison of sentence processing theories on naturalistic text. We find that both theories can explain some of the variance in the eye-movement data – while structural Surprisal is a significant positive predictor across the complete data set, DLT integration cost correctly predicts variance on verbs (for which it makes the bulk of its predictions). In addition, we show that the two theories capture different aspects of sentence processing: their predictions are uncorrelated. While the finding that Surprisal and DLT predictions are uncorrelated is not surprising, it nicely supports experimental case studies that show that DLT and Surprisal can explain different processing difficulty phenomena.

[1] These and other theories of sentence processing are explained in more detail in Chapter 2. A comparative evaluation to the theory developed in this work is provided at the end of Chapter 9.

4 Proposal of a new Sentence Processing Theory

Chapter 6 proposes a new theory of sentence processing, which is psycholinguistically motivated in that it models *strict incrementality* (this means that a word is eagerly integrated with earlier structure as soon as it is perceived) and an explicit *prediction* and *verification* process, as well as *memory decay* and *parallel processing*. Evidence for prediction comes for example from the finding that people are able to anticipate the argument of a verb (as measured through increased fixations on the argument in a visual world paradigm even before this argument is vocalised, (Kamide et al., 2003)). Additional evidence for prediction comes from experiments where N400 effects are observed when the form of the determiner does not match the most strongly anticipated noun (van Berkum et al., 2005), and the processing of *either..or* constructions where the word *or* and a following noun phrase were read faster in contexts that included the word *either* (Staub and Clifton, 2006). This effect is explained as the word *or* and the second conjunct being predicted when processing the word *either*. Interestingly, the assumption of strict incrementality, where all words always have to be connected under a single root node, automatically leads to predictions (e.g., the structure of an upcoming head has to be predicted in order to connect two seen dependents).

In addition to the fundamental assumptions of strict incrementality, prediction and verification, the proposed sentence processing theory, which we will refer to as "Prediction Theory" in the following, unifies the complementary aspects of Surprisal and DLT into a single theory. Prediction Theory has two mechanisms that account for processing difficulty: The concept of Surprisal is used to quantify the difficulty of the parser in terms of updating its representation of the analyses as the sentence unfolds. In addition, difficulty can arise at integration time, when validating previously predicted structures against what is actually encountered. The amount of difficulty generated in verification depends on (a) how difficult the prediction was and (b) on how recently the prediction was made: if the prediction has decayed a lot, more difficulty arises than when a structure was predicted very recently. The model therefore needs to keep track of when each syntactic node was predicted, which is realised through *time stamps* on the nodes. This verification process thus causes difficulty based on a memory retrieval process for retrieving and integrating newly encountered structure with previously predicted structure. These two types of processing difficulty, updating one's representations of the sentence on the one hand, and memory retrieval and integration on the other hand thus model theoretically different aspects of human sentence processing.

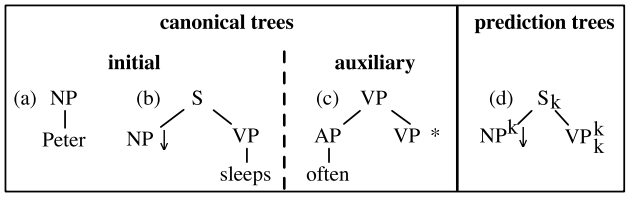
5 Developing a cognitively plausible grammar formalism for Prediction Theory

In order to adequately implement Prediction theory, it is necessary to choose a parser (and thereby a grammar formalism) that supports strict incrementality, prediction and verification. Most incremental parsers are however not strictly incremental, but instead maintain unconnected partial structures on a stack. The existing parser that satisfies the requirements of our theory best is Roark's (2001) top-down PCFG parser, as it is strictly incremental, scales up to broad-coverage parsing and uses a generative model (which is useful as Surprisal can be directly calculated from such an incremental generative model). Drawbacks of the Roark (2001) parser are that it uses a top-down parsing strategy, which has been argued to be less cognitively plausible than a left corner arc-eager parsing strategy (Abney and Johnson, 1991) and that using a context-free grammar is less cognitively plausible than using a grammar formalism that is mildly context-free. We therefore decided to develop our own cognitively more plausible parser in order to adequately implement Prediction theory. The last part of chapter 6 discusses the suitability of alternative grammar formalisms and concludes that a strictly incremental version of Tree-adjoining Grammar (TAG) would most adequately reflect the stated mechanisms of the processing theory. Tree-adjoining grammar is mildly context-sensitive and it supports an extended domain of locality (Joshi, 2004), which is more powerful than e.g. a context free grammar in locally describing the relationships between words. The new incremental version of the TAG formalism, called Psycholinguistically motivated TAG (PLTAG) is introduced in Chapter 7. We motivate the development of this incremental variant by first showing that standard TAG cannot incrementally derive even very simple sentences such as "Peter often sleeps." (not shown here), then formally define PLTAG, and finally demonstrate the equivalence of TAG and PLTAG.

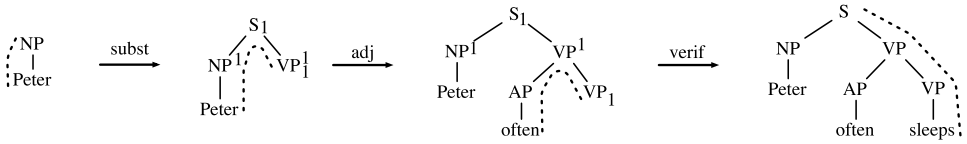
Lexicalized Tree Adjoining Grammar (LTAG, Joshi et al. 1975) is a grammar formalism whose lexicon consists of *elementary trees*, each of which is anchored by a lexical head. Grammatical derivations are built from these elementary trees by two tree-combining operations, *substitution* and *adjunction*. PLTAG introduces *prediction trees* as a second type of lexicon entry in addition to the usual elementary trees of LTAG (which we call *canonical trees*, see Figure 1(a)). Prediction trees are (not necessarily lexicalized) elementary trees in which each node carries one or two *markers* indicating that this node is only being hypothesised by the parser during the course of an incremental derivation. Prediction trees can be substituted and adjoined in exactly the same way as canonical trees; the markers of a prediction tree are instantiated with fresh symbols by these operations, so we can always tell

from a derived tree which nodes came from the same prediction tree.

Markers are eliminated from a partial derived tree through a new operation called *verification*. The verification operation validates the nodes introduced by a prediction tree in an earlier derivation step by matching them with the nodes of a canonical elementary tree. We refer to an elementary tree that verifies an earlier prediction as the *verification tree*. The verification operation assumes that the verification tree ε is *compatible* with all the nodes carrying a certain marker *k*; this means that the verification tree contains all nodes marked with *k* in exactly the same position as they were in the original prediction tree. Crucially, ε is allowed to contain nodes to the right of its spine[2] that do not occur in the partial derived tree as nodes with marker *k* (but not nodes to the left of the spine). This reflects the asymmetry of incrementality. After verification, the markers on all verified nodes are removed. A PLTAG derivation including a substitution, an adjunction and a verification operation for the sentence "Peter often sleeps" is shown in Figure 1.



(a) A grammar for PLTAG



(b) A derivation in PLTAG using the trees from the example grammar in subfigure (a); the dotted line indicates which part of the derived prefix tree is relevant for the next operation.

Figure 1: The PLTAG formalism.

A *PLTAG derivation* is always incremental. It starts with the trees of the first input word, and then applies substitution, adjunction, and verification operations as follows: if the first i leaves of the derived tree at some point in the derivation are the words $w_1 \ldots w_i$ and the next derivation step is a substitution, adjunction, or verification with a canonical tree, then the anchor of this tree must be w_{i+1}. We call a derivation of a sentence $w_1 \ldots w_n$ *complete* if $i = n$, the derived tree contains no more substitution nodes, foot nodes or prediction markers, and the root symbol of the derived tree is S.

[2] A tree's spine is the path from the root to its anchor leaf. This is usually the head of that tree.

6 A Parser for Implementing Prediction Theory

Chapter 8 describes the complete development of the PLTAG probabilistic parser, beginning with the creation of a PLTAG treebank, which is obtained by automatically converting the standard Penn Treebank into PLTAG format, and the extraction of the PLTAG lexicon from the converted treebank. The PLTAG parsing algorithm generates multiple analyses for a string in parallel and uses an eager left-corner parsing strategy. Section 8.3 formally defines the parser operations and proves that the parser always generates valid PLTAG derivations. In order to make the parser efficient enough for broad-coverage parsing, a number of optimisations are necessary for the implementation. These include storing alternative analyses in a chart in order to avoid executing equivalent operations multiple times, using a beam to only follow up on the most probable analyses, introducing a supertagger to choose the most probable predictions to make after processing each word and restricting the parser in terms of the number of predictions it can make at once. These optimisations cause the parser to be incomplete (but this is to some extent the case for all parsers that do beam search, i.e. all tractable parsers). A probability model for the parser is developed in Section 8.5, and the parser is evaluated on a standard test set (section 23 of the Penn Treebank) in Section 8.6. Evaluation results show that the parser achieves broad coverage and a suitable accuracy (approaching the performance of non-incremental TAG parsers) for evaluating the sentence processing theory based on this parser on a broad-coverage corpus. The final section of chapter 8 describes the linking theory which defines how the parsing process translates into processing difficulty estimates for each word.

The strictly incremental parser developed here is also of potential interest to other areas of computational linguistics. Strictly incremental parsers can find application in speech-to-speech translation or timely reactions to ongoing speech in agents.

7 Evaluation of Prediction Theory

Chapter 9 evaluates the psycholinguistic aspects of Prediction Theory by testing it both on a selection of established sentence processing phenomena and on the Dundee eye-tracking corpus. The predictions of the implemented sentence processing model are evaluated and discussed with respect to nine different psycholinguistic case studies. The first case study concerns the well-known SRC / ORC asymmetry, which describes the phenomenon that subject relative clauses like *...who attacked the senator...* are easier and faster to process than object relative clauses like *...who the senator attacked...*

A recent study by Staub (2010) investigated where exactly the difficulty occurs. Staub (2010) found that go-past reading times are longer in the ORC condition both on the embedded verb in the relative clause (*attacked*) and on the embedded NP (*the senator*), see the top left bar chart in Figure 2. In order to evaluate the predictions of Surprisal, DLT and our Prediction theory on the experimental data, all three models were run on the experimental materials used in Staub (2010). As can be seen in the right column of Figure 2, Surprisal only predicts a significant difference in processing difficulty on the onset of the noun phrase, while DLT only predicts the increased difficulty on the embedded verb. Prediction Theory however predicts increased difficulty in the ORC condition both on the onset of the NP region and on the embedded verb, see the bottom left bar graph in Figure 2. The observed longer reading times on the noun in the empirical data, which is not predicted by any model, can be explained as a spill-over effect from the difficulty incurred at the onset of the noun phrase. The determiner is very rarely fixated at all, so any difficulty occurring there only becomes apparent at the next fixation, which is most often the following noun.

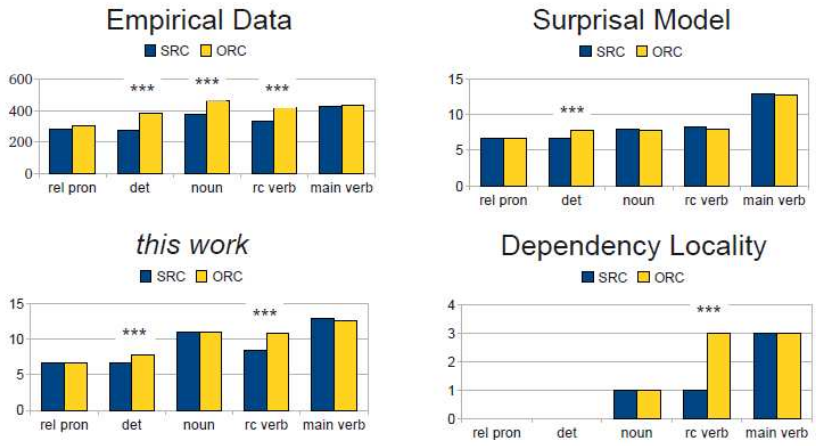


Figure 2: The Relative Clause Asymmetry; the top left bar chart reports go-past reading times in msec, while the other graphs report the predicted processing difficulty.

In addition to the experimental materials, we also evaluated our Prediction theory on the relative clauses extracted from the Dundee Corpus, like done for Surprisal and DLT in chapter 4. The predictions by our theory correctly account for the observed reading time on the Dundee corpus relative clauses, and turn out to predict the data from the verb region of these naturalistic relative clauses better

than either Surprisal or DLT integration cost.

A second case study tested the effect of the presence of the word *either* on the later occurrence of the word *or* and the following constituent. Surprisal, DLT and Prediction Theory were run on materials from the experiment by Staub and Clifton (2006). Both Surprisal and our theory correctly predicted a facilitation at the word *or* and the following constituent, while DLT did not predict a significant difference. The thesis discusses a range of other psycholinguistic effects and shows that Prediction theory can account for more of the effects simultaneously than either DLT or Surprisal (or any other current sentence processing theory).

The second part of the evaluation chapter evaluates Prediction theory on the reading times from the Dundee corpus. This broad-coverage study complements the experimental results, which only focus on very specific psycholinguistic phenomena, by testing whether a processing theory has explanatory power also for the reading times observed on the wide range of structures present in naturalistic text. We parsed the Dundee corpus with our incremental PLTAG parser and automatically calculated difficulty predictions for each word, just as done for Surprisal and DLT integration cost in Chapter 5. As one would expect from the design of Prediction Theory, we can show that its difficulty predictions are correlated with both lexical surprisal and DLT integration cost. Prediction Theory processing difficulty estimates turn out to be a significant positive predictor for first fixation, first pass and total reading times on the naturalistic text. In a comparative evaluation focussing on the explanatory power of the alternative sentence processing theories, Prediction Theory is shown to explain a larger proportion of the variance in the reading times than either DLT integration cost or Surprisal.

In conclusion, we find a wide range of empirical support for the PLTAG-based theory of prediction and verification in human sentence processing, and show that it has larger explanatory power also on general, naturalistic text than previous theories of sentence processing.

8 Conclusions and Directions for Future Research

The most significant contributions of this thesis are the demonstration of the usefulness of evaluating sentence processing theories on broad-coverage, naturalistic text in addition to standard lab experiment materials, and the design, full implementation and evaluation of Prediction Theory.

This thesis is interdisciplinary in that having a broad-coverage model of human sentence processing that accurately predicts processing difficulty on the syntactic level is not only of theoretic interest to psycholinguistics, but also highly relevant for researchers in computational linguistics. Such a fully automatic system for de-

termining the syntactic processing difficulty incurred when reading a text can contribute to automatically assessing the difficulty of a text (such systems can be used in automatic readability assessments) and optimisation of machine-generated text or speech (for example in symmetrisation, translation, tutoring systems and general dialogue systems). Finally, the development of PLTAG, a psycholinguistically motivated version of Tree-Adjoining Grammar that supports strict incrementality as well as explicit mechanisms for prediction and verification constitutes a significant contribution to the field of linguistics.

The research conducted for this thesis also leads to a number of future research questions, outlined in Section 10.2 of the thesis. The most interesting directions in further developing the Prediction Theory model are to enable it to also account for language acquisition effects by gradually acquiring its lexicon and probability model over time, and also introduce a dynamical update to the language model, thus being able to model short term priming effects as well as long term learning. A further important shortcoming of the current model is its restriction to the syntax level. In future work, it is planned to extend this model to the semantic and discourse levels, and thus account for a larger proportion of the processing difficulties that humans encounter when comprehending language. This will also allow us to account for a larger amount of the variance in reading time data.

References

Abney, S. P. and Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*.

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

Hale, J. (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*, volume 2, pages 159–166, Pittsburgh, PA. Association for Computational Linguistics.

Joshi, A., Levy, L., and Takahashi, M. (1975). Tree adjunct grammars. *Journal of the Computer and System Sciences*, 10.

Joshi, A. K. (2004). Domains of locality. *Data Knowledge Engineering*, 50(3):277 – 289. Special jubilee issue: DKE 50.

Kamide, Y., Scheepers, C., and Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from german and english. *Journal of Psycholinguistic Research*, 32:37–55.

King, J. and Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30:580–602.

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

Lin, D. (1998). An information-theoretic definition of similarity. In Shavlik, J. W., editor, *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, Madison, WI. Morgan Kaufmann.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116:71–86.

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

van Berkum, J. J., Brown, C., Zwitserlood, P., V. Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31:443–467.