Emulation of Human Sentence Processing using an Automatic Dependency Shift-Reduce Parser

Atanas Chanev

Logical Factor Bulgaria, Yakubitsa Street 2a, 1164 Sofia, Bulgaria

Abstract. The methods of NLP and Cognitive Science can complement each other for the design of better models of the human sentence processing mechanism, on the one hand, and the development of better natural language parsers, on the other. In this paper, we show the performance of an automatic parser consistent with the architecture of the human parser of [2] on various human sentence processing experimental materials. Moreover, we use a linking hypothesis based on the concept of surprisal [9] to explain human reaction time patterns. Although our results are generally not consistent with the human performance, our emulations contribute to understanding the architecture of the human parser and its disambiguation strategies better. We also suggest that these strategies may possibly be used for improving the performance of automatic parsers.

1 Introduction

Some of the models of the human sentence processing mechanism reported in the recent years are capable of achieving wide coverage on random corpora (e.g. [7]). Most of these models are based on natural language parsing algorithms. On the other hand, automatic parsers can benefit from knowledge about the way humans process sentences in natural languages (e.g. see [2] for examples and discussion).

The process of selection and extension of a natural language parser to a model of the human sentence processing mechanism has been reported in [2]. They show that the porting can be done in three stages, as illustrated below: preparing a list of constraints for the model based on general knowledge about the human parser, as well as evidence from experiments with human subjects; design of the architecture and association of a linking hypothesis which should be capable of emulating and explaining reaction times of humans in sentence processing experiments.

$$Constraints \rightarrow Architecture \rightarrow LinkingHypothesis$$

In this paper, we report several emulations of the human sentence processing mechanism using sentences from [9]. We have used an automatic dependency parser, [1] which is compatible with the model of the human parser of [2]. However, we have used a new linking hypothesis to explain the relationship between the architecture of the parser and the performance of human subjects in sentence processing experiments. This linking hypothesis is based on the concept of surprisal, [9].

Chanev A. (2008).
Emulation of Human Sentence Processing using an Automatic Dependency Shift-Reduce Parser.
In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, pages 137-146
DOI: 10.5220/0001732001370146

We emulated properly human difficulty for only one of five sets of experimental materials. We have performed error analysis in the remaining cases to discover the reasons for the performance of our models. We have suggested three ways to improve our models in order to emulate human sentence processing better.

The paper is structured as follows: In Section 2 we review the model of [2]. Then, in Section 3, we present our linking hypothesis for explaining human difficulty patterns. We describe our experimental settings in Section 4. Then, in Section 5, we report our results. We conclude in Section 6 and list our future plans in Section 7.

2 The Model of (Chanev, 2007) Revisited

(Chanev, 2007), [2] argue for the psychological plausibility of the class of deterministic dependency shift-reduce parsers. They propose an architecture of the human sentence processing mechanism. Compared to connectionist models, it is more robust and has a wider coverage, whereas compared to other broad coverage models of the human parser, it is more detailed than e.g. [9], and its parsing algorithms are more incremental than e.g. the top-down algorithm used in [7][1].

We use an automatic parser from the class of dependency shift-reduce parsers in our experiments. The basic features of [2] are described below. They include the constraints satisfied by the model, the architecture and a linking hypothesis.

The model of [2] satisfies the following constraints:

General Constraints. Wide coverage, high accuracy, robustness and multilinguality.

Architectural Constraints. Incrementality and non-projectivity[2].

Informational Constraints. Lexical frequency, discourse context, semantic plausibility, prosodic breaks and syntactic preferences.

The model of [2] uses Dependency Grammar, e.g. [11]. Thus, it recognizes sentence structure as a set of binary head-dependent relations. In Figure 1, we show the dependency structure of a simple sentence. Dependency grammar can be considered a good choice for a model of the human parser, e.g. because non-projective relations can be encoded in the syntactic tree easily.

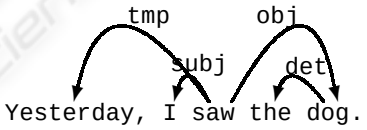

Fig. 1. The dependency structure of a simple sentence.

[1] The interested reader is referred to [2] for a detailed comparison of the dependency shift-reduce parsing architecture with other models of the human parser.

[2] Non-projectivity is the ability of a parser to process non-projective grammatical relations, e.g. the one in the sentence *"I saw the dog yesterday with the red nose."* between the head of the noun phrase *The dog* and the head of the prepositional phrase *with the red nose*.

The architecture of [2] is serial (i.e. it is not parallel). A scheme is shown in Figure 2. The following components can be distinguished:

Stack – for storing partially processed word tokens;

Input Memory – for tokens that have not been yet integrated to a temporary sentence structure or stored in the stack;

Parsing Actions – for pushing tokens into the stack and popping them out, as well as to build dependency relations between tokens;

Memory for Processed Tokens – used for tokens that have been popped from the stack or removed from the input memory;

Classifier – used to learn the sequence of parsing actions needed to build the dependency tree of the sentence.

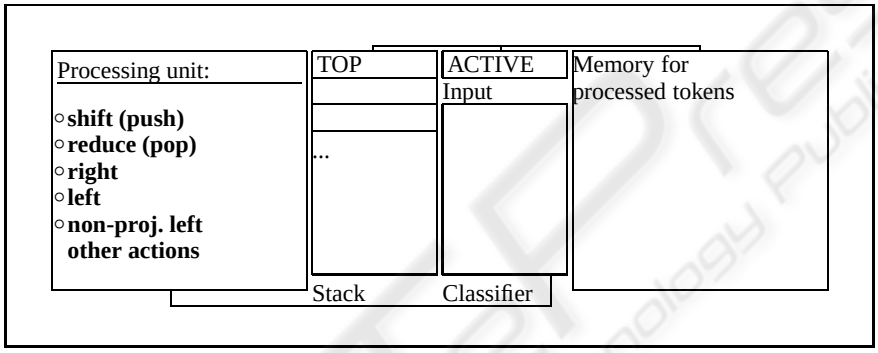

Fig. 2. The model of the human sentence parsing mechanism of [2].

The classifier component of the model consists of three functional modules: a database of language experience; a learning algorithm and a feature model. The latter specifies the configuration of features of certain word tokens in the sentence to be learned for determining next parsing actions.

One of the linking hypotheses of [2] is influenced by the Surprisal model [7, 9]. The study of [7] was the first to measure the surprisal associated with the integration of each word into a temporary syntactic structure of the sentence. They calculate the surprisal using prefix probabilities [14]. Then, the model has been generalized in [9] for arbitrary grammars and parsing algorithms deriving the formula for surprisal on information theoretical grounds. Both of the models assume parallel activation of syntactic structures.

In [2] difficulty is measured similarly to the way it is measured in the Surprisal model. However, since the main component of their architecture is a serial dependency shift-reduce parser, the likelihood, as assigned by the classifier, is measured instead of prefix probabilities. It is calculated with respect to the particular learning algorithm used by the classifier and over parsing actions rather than syntactic sub-trees.

3 Our Linking Hypothesis

Our linking hypothesis is in the spirit of the Surprisal model. It is naturally implemented in the automatic parser that we use, DeSR[3], [1] and is defined in terms of the Average Perceptron learning algorithm, [4] as implemented in the parser.

We had to adapt the Surprisal model which assumed a parallel architecture, to DeSR which is a serial parser. However, multiple activations of different syntactic interpretations, as in the Surprisal model, can still be emulated in DeSR through feature models that avoid learning syntactic dependencies explicitly. In addition, DeSR can measure the human difficulty, using spans of the sentence that does not necessarily begin with its first word, unlike [9]. This makes our model more flexible than the Surprisal theory while still possessing its basic characteristics.

The Average Perceptron is a multi class perceptron [4]. Each of the classes is a parsing action, e.g. shift, right (subject), right (determiner), left (object) etc. At each step, the most likely action is executed. In order to measure the difficulty associated with the integration of a token, we use the likelihood of the integrating parsing action. Thus, the higher the likelihood is, the lower the human difficulty would be.

It must be noted that a word token can be integrated to the temporary syntactic structure of the sentence as a syntactic dependent exactly once, and as a syntactic head, zero or more times. In the cases where the integration of the word is done through more than one dependency relations, the difficulty is measured as the sum of the difficulties of building all the relations between the word, its partially processed dependents, if any, and its partially processed head word.

4 Experimental Settings

Corpus. We used the training set of the dependency version of the Penn treebank [10] as used in the CoNLL 2007 shared task on dependency parsing [13]. In this format, the treebank is annotated with part-of-speech tags and dependency syntax. Moreover, we used the supersense tagger of [3] to annotate the texts in the treebank with semantic WordNet classes. We merged all the information into one resource and used it to train our models. We annotated our test set with part-of-speech information, using the SVM-Tagger[4] [5] and corrected the errors manually. Then, we used the supersense tagger for annotating it with semantic classes. Finally we merged all the information.

Feature Models. We started our experiments with the best model for English that was in the package of the DeSR parser but removed some of the features that we had found implausible for a model of the human parser. These include the properties of tokens that are too far ahead in the sentence. We also added semantic features and used a first order Average Perceptron. We trained two models, *Base* and *Syntactic*, because we wanted to distinguish between a model that uses information about a particular syntactic structure for making parsing decisions (*Syntactic*) from a model that does not use such information (*Base*). *Base* should score similarly to the surprisal model due to measuring

[3] Freely available from http://sourceforge.net/projects/desr/
[4] http://www.lsi.upc.es/∼nlp/SVMTool/

the total likelihood of a string instead of the likelihood of a specific structure associated to a string.

We show the model that has syntactic features in Table 1. We use the notation of DeSR. The tokens in the stack are encoded with negative numbers where -1 is the token that is on the top of the stack. The tokens in the input are encoded using 0 and positive numbers where 0 is the first token in the input.

Table 1. Our syntactic feature model.

Feature	tokens	Feature	tokens	Feature	tokens	Feature	tokens
LexFeatures	-2 -1 0 1	LexPrev	0	LexSucc	-1	SemFeatures	-2 -1 0 1
SemLeftChild	0	SemRightChild	-1	SemPrev	0	SemSucc	-1
PosFeatures	-2 -1 0 1	PosLeftChild	-1 0	PosRightChild	-1 0	PosPrev	0
PosSucc	-1	CPosFeatures	-1 0 1	DepLeftChild	-1 0	DepRightChild	-1

Parsing Accuracy. We show the accuracy of the parser in Table 2. High accuracy is an important pre-requisite for the proper emulation of human reaction times because the patterns obtained from integrating inaccurate dependency relations would be very different from human difficulty patterns, in the general case.

Table 2. The accuracy of our models.

Model	Labeled Attachment Score	Unlabeled Attachment Score
Base	77.6%	79%
Syntactic	77.5%	78.9%

We solve the problem of erroneous disambiguation by 'forcing' the parser to execute certain actions which are less likely than those which it otherwise would have executed, if guided by the classifier. Still, if the parser has a very low accuracy, that would 'distort' the likelihood of parsing actions and the predicted difficulty, respectively.

5 Emulation of Experiments with Human Subjects

We have emulated human behavior on sentences from five experiments. They have been selected among the experiments used by [9] to evaluate the Surprisal theory[5]. We report the performance of our models on only one sentence from each condition of each experiment. We do not need to average over many experimental items, because the models have sufficient linguistic knowledge to make parsing decisions in a plausible way.

The human difficulty patterns for certain regions in the sentences of two experiments have been consistent with the Surprisal theory. In the other three experiments, difficulty patterns for certain regions of the sentences are not consistent with the Surprisal theory. We have tried to emulate human behavior for all the sentences using our models.

[5] Our models have been tested on all the experimental materials used in [9]. We report only five experiments due to the lack of space. However, they are sufficient to illustrate the most important aspects of using our models on the data of [9].

Experiment 1. The first experiment is reported by [8]. It investigates the relevance of surprisal in three sentences with a different number of words between the subject modified by a relative clause and the main verb, for the difficulty observed at the main verb. The sentences are given below:

1. The Player [that the coach met at 8 o'clock] bought the house...
2. The Player [that the coach met by the river at 8 oclock] bought the house...
3. The Player [that the coach met near the gym by the river at 8 oclock] bought the house...

The main verb of sentence 3 has been read faster than the one of sentence 2 which has been read faster than the one of sentence 1. The prediction pattern of the surprisal theory is the same as the human reaction times (or difficulty) pattern. The explanation of the Surprisal theory is that with the increase of the number of words in the relative clause, the expectation for the main verb increases.

However, we report a different pattern. The *Base* model predicted the same difficulty at the main verb for all the sentences. The likelihood numbers of the Average Perceptron for the integration of the main verb are the same: 119 for attaching the subject to the main verb and 166 for the recognition of the main verb as the root of the sentence. On the other hand, the *Syntactic* model predicted a difficulty for sentence 2 (likelihood of 166 for the subject relation and 194 for the root relation), if compared to sentences 1 and 3 (146 for the subject relation and 194 for the root relation).

Our models cannot account for surprisal effects, because at the time of parsing the features of no previous tokens that are not in the stack except for the adjacent left token are taken into consideration. Thus, the tokens responsible for the parsing decision are the same for the three sentences, for the *Base* model or almost the same for the *Syntactic* model. The way to account properly for the differences in the sentences is to include more past tokens in the feature models or to include features for distance.

Experiment 2. Experiment 2 uses materials from [16]. In their experiments it has been shown that an unresolved ambiguity can facilitate comprehension. The sentences are given below:

1. The daughter of the colonel who shot herself on the balcony had been very depressed.
2. The daughter of the colonel who shot himself on the balcony had been very depressed.
3. The son of the colonel who shot himself on the balcony had been very depressed.

In [16] they have measured the difficulty of integrating the relative pronoun to the temporary syntactic structure. They have shown that the difficulty of integrating the relative pronoun in sentences 1 and 2 is greater than the one of integrating it in sentence 3. The Surprisal model predicts this pattern because the integration in the ambiguous case would have a probability that is the sum of the probabilities of the two different interpretations. The greater probability would mean a smaller surprisal and a smaller difficulty, respectively.

The *Base* model predicted a likelihood of 116 for sentence 1; 107, for sentence 2 and 100 or 107, for sentence 3, depending on the syntactic interpretation. On the other hand, the *Syntactic* model predicted a likelihood of 132 for sentence 1; 126, for sentence 2 and 120 or 126, for sentence 3 depending on the syntactic interpretation. Taking the sum of the likelihoods of the two interpretations of sentence 3, both of our models would emulate the human pattern.

The only demerit of that assumption is that our model is strictly serial. Summing likelihoods of different attachments would assume parallel processing at least at some point in the sentence. However, there might be an alternative interpretation of the results. The DeSR parser cannot distinguish between the ambiguous situation and the situation where the relative pronoun is attached low (i.e. in sentence 2). The solution is to use differently defined parsing actions, e.g. the actions of Maltparser [12]. This would also make the parsing algorithm more incremental.

Experiment 3. In Experiment 3 we use materials from [6]. In the sentences below, matrix verbs from subject extracted relative clauses have been easier to process by humans than those from object extracted relative clauses. It must be noted that the surprisal model cannot explain the observed difficulty pattern.

1. The reporter who sent the photographer to the editor hoped for a good story.
2. The reporter who the photographer sent to the editor hoped for a good story.

There are two dependency relations that are to be built for the integration of *'sent'* to the temporary syntactic structure of the sentence with the subject extracted relative clause. One of them is the subject relation between *'who'* and the matrix verb of the relative clause. The other is the modifier relation between *'reporter'* and the matrix verb. The dependency relations to be built for the integration of *'sent'* in the object extracted relative clause are: the subject relation between *'photographer'* and *'sent'*; the modifier relation between *'reporter'* and *'sent'* and the object relation between *'who'* and *'sent'*. The difficulty at the integration of *'sent'* would depend on the total likelihood of all the syntactic relations to be built between the matrix verb of the relative clause and other tokens.

The *Base* model integrates *'sent'* in sentence 1 with a likelihood of 268 (133 for the subject relation and 135 for the modifier relation). It integrates *'sent'* in sentence 2 with a likelihood of 144 (4 for the subject relation, 116 for the modifier relation and 24 for the object relation). The pattern is as predicted because the more likely integration causes less difficulty. The *Syntactic* model shows a similar pattern. For sentence 1, the matrix verb of the relative clause is integrated with a likelihood of 300 (146 for the subject relation and 154 for the modifier relation). In sentence 2 the likelihood of the integration of *'sent'* is 220 (38 for the subject relation, 159 for the modifier relation and 23 for the object relation). The pattern is again as predicted.

Experiment 4. In experiment 4 we use materials from [6]. There are three sentences with an object extracted relative clause modifying the subject of the sentence. The subject in the relative clause is with varying length. The sentences are given below:

1. The administrator who the nurse supervised scolded the medic ...
2. The administrator who the nurse from the clinic supervised scolded the medic ...
3. The administrator who the nurse who was from the clinic supervised scolded the medic ...

In [6] it is reported that the difficulty associated with the integration of the verb of the relative clause increases with the increase of the distance to the subject of the relative clause. This means that the verb of the relative clause of sentence 1 will be less difficult to integrate than the one in sentence 2 which will be less difficult to integrate

than the one in sentence 3. It must be noted that the Surprisal model cannot explain the observed difficulty pattern.

There are three syntactic relations to be built in order to integrate *'supervised'* into the temporary syntactic structure of the sentence. They are: the subject relation between *'nurse'* and *'supervised'*; the modifier relation between *'administrator'* and *'supervised'*; and the object relation between *'who'* and *'supervised'*.

In this experiment both of our models predict what the surprisal model would. That is, with the increase of distance, difficulty will not increase but decrease. The *Base* model assigns likelihoods of 258 (117 for the subject relation; 51 for the object relation and 90 for the modifier relation) for sentence 1; 265 (124 for the subject relation; 51 for the object relation and 90 for the modifier relation) for sentence 2 and 280 (139 for the subject relation; 51 for the object relation and 90 for the modifier relation) for sentence 3.

The patterns of the *Syntactic* model are similar to those of the *Base* model. The likelihood of the integration of the verb of the relative clause is 301 (129 for the subject relation; 55 for the object relation and 117 for the modifier relation) for sentence 1; 322 (152 for the subject relation; 55 for the object relation and 115 for the modifier relation) for sentence 2 and 338 (168 for the subject relation; 55 for the object relation and 115 for the modifier relation) for sentence 3.

The possible reason for our results is the lack of features of past tokens that are not in the stack of the parser. We should also mention that distance can be included as a feature in a parsing model and used in the learning phase.

Experiment 5. In experiment 5 we have used materials from [15]. They have shown that reduced relative clauses with non-subject context where the modifying verb has different forms for past tense and past participle, are easier to understand than reduced relative clauses with non-subject context where the modifying verb is ambiguous. It must be noted that the surprisal model cannot explain the observed difficulty pattern. The sentences are given below:

1. The coach smiled at the player thrown the frisbee.
2. The coach smiled at the player tossed the frisbee.

In [15] it has been shown that the integration of *'thrown'* into the syntactic structure of sentence 1 is easier than the integration of *'tossed'* into the syntactic structure of sentence 2. In both of the sentences, the verb in the relative clause is attached to the preceding noun with a modifier relation.

None of our models shows any clear pattern. The *Base* model attaches *'thrown'* to *'player'* with a likelihood of 125 and *'tossed'* to *'player'* with a likelihood of 126. These numbers for the *Syntactic* model are 145 and 144, respectively. The major issue in this experiment is the use of the past participle part-of-speech tag for the verb *'tossed'*. The use of this tag assumes that the main-verb/reduced relative clause ambiguity has been resolved and it should not have been resolved at the time of integration.

There are two ways to overcome this demerit of the parser: to use a single tag for the verbs in past tense and participle form or to integrate part-of-speech tagging into parsing vertically. Both of these solutions have their demerits and bottlenecks. For example, the implementation of the former would prevent the parser from using valuable information

in disambiguation. The implementation of the latter would put too much weight on the knowledge-poor part-of-speech component to take important parsing decisions.

6 Conclusions

In this paper we have used two models to parse experimental materials from a number of studies. Our models have predicted the pattern of human difficulty for only one of five experiments. The reasons for the performance of the models in the other four experiments can be classified as follows:

In experiment 1 and experiment 4, the reason is the lack of features of previous tokens that are not in the stack. It may also be useful to use the number of words between a head and its dependent as a feature in the feature model. It must be noted that the Surprisal model can explain human difficulty patterns in experiment 1 but it cannot explain them in experiment 4. We believe that there is one and the same reason for the performance of our models on the materials from both experiment 1 and experiment 4.

The reason for the performance of our model on the materials from experiment 2 is the definition of the parsing actions of the parser that we use. This way, our models cannot distinguish between an unresolved ambiguity case and one of the unambiguous cases at the time of integration. One way to address this demerit is to use differently defined parsing actions such as the actions of another dependency shift-reduce parser, Maltparser [12]. This would allow processing in a more incremental way, as well.

For experiment 7, part-of-speech tagging and parsing should be integrated in a reasonable way, in order for the experiment to be plausibly conducted. A possible integration should result in a main verb/reduced relative clause disambiguation performed by both the part-of-speech tagger and the parser, at the same time.

Taking into consideration our emulation results, we have identified three ways to make our models of the human parser more plausible. They are: a change in the feature model for training; using parsing actions that would make processing more incremental and the integration of part-of-speech tagging and parsing for joint disambiguation of the main-verb/reduced relative clause ambiguity. We expect that using these techniques, it could be possible to increase the accuracy of the parser, as well.

7 Future Work

For future work, we intend to change the feature model of the parser incorporating features of previous tokens and possibly, a feature for the distance between the heads and their dependents. We intend to use a parser with differently defined actions in order to emulate properly the human difficulty pattern for experiments in which ambiguity facilitates processing. We consider to initiate the integration of part-of-speech tagging into the parsing process. We will also explore the way that these improvements would affect ambiguity resolution and the accuracy of the parser.

Acknowledgements

Most of the work has been completed while the author was still affiliated to University of Trento and FBK-irst, Povo-Trento. We would like to acknowledge Paolo Bouquet, Alberto Lavelli and Edward Gibson for valuable discussions on the use of the dependency shift-reduce parser as a model of the human sentence processing mechanism. We thank two anonymous reviewers for their comments on a previous draft of this paper.

References

1. Attardi, G., Dell'Orletta, F., Simi, M., Chanev, A. & Ciaramita, M. (2007). Multilingual Dependency Parsing and Domain Adaptation using DeSR. In Proc. of the shared task session of EMNLP-CoNLL 2007, Prague
2. Chanev, A. (2007). Investigating the relationship between automatic parsing and human sentence processing. PhD thesis, University of Trento
3. Ciaramita, M., & Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In Proc. of EMNLP 2006.
4. Crammer, K. & Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. Journal of Machine Learning Research, 3, 951–991.
5. Giménez, J., & Márquez , L. (2004). SVMTool: A general POS tagger generator based on Support Vector Machines. Proc. of LREC'04, Lisbon, Portugal.
6. Grodner, D. & Gibson, E. (2005). Some consequences of the serial nature of linguistic input. Cognitive Science, 29(2), 261–290.
7. Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proc. of the 2nd Meeting of NAACL.
8. Jaeger, F., Fedorenko, E., & Gibson, E. (2005). Dissociation between production and comprehension complexity. Poster Presentation at the 18th CUNY Sentence Processing Conference, University of Arizona.
9. Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106 (3), 1126–1177.
10. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19 (2).
11. Melčuk, I. (1988). Dependency syntax: Theory and practice. State University of New York Press.
12. Nivre, J. (2006). Inductive Dependency Parsing. Springer.
13. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S. & Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In Proc. of the shared task session of EMNLP-CoNLL 2007, Prague
14. Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. Computational Linguistics, 21(2), 165–201.
15. Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. Journal of Memory and Language, 50(4), 355–370.
16. Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. Journal of Memory and Language, 39, 558–592.