A Probabilistic Parser as a Model of Global Processing Difficulty

Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK

Abstract

We present a model of global processing difficulty in human parsing. This model is based on a probabilistic context-free grammar and is trained on a realistic corpus sample. It achieves broad coverage and good parsing accuracy on unseen text, and its predictions are significantly correlated with experimental data on word order preferences in German. The model makes predictions about the differential behavior of verb final and verb initial sentences and provides evidence for the importance of lexical information in sentence processing.

Introduction

The human sentence processor is constantly confronted with ambiguous input, i.e., with linguistic material that is compatible with more than one syntactic analysis. The question of how such ambiguities are resolved has generated considerable debate in the psycholinguistic literature, and a variety of approaches have been proposed to address this question (see Crocker, 1999 for an overview).

One of these approaches is the *Tuning Hypothesis* (Mitchell, Cuetos, Corley, & Brysbaert, 1996), which states that the sentence processor extracts frequency information from its environment. In the case of ambiguity, the processor adopts the most frequent structure. This predicts that more frequent structures are easier to process than less frequent ones, as the processor is more likely to have encountered them before, and can choose the correct analysis. A number of experimental studies confirm this prediction for PP attachment and relative clause attachment (e.g., Brysbaert & Mitchell, 1996).

The aim of the present paper is to generalize the Tuning Hypothesis from *local processing difficulty* (as it occurs with attachment ambiguities) to *global processing difficulty*. By this we mean processing difficulty that persists even when the whole sentence has been read in by the parser, and a unique reading should be available. Standard examples include center embedding constructions and constructions that induce 'strong' garden paths, i.e., garden paths from which the parser fails to recover. The Tuning Hypothesis predicts a link with frequency for these cases: if the globally correct analysis of a sentence is infrequent, then this sentence will lead to strong processing difficulty.

In this paper, we test prediction with respect to word order variation in German, a phenomenon for which global processing difficulty can be observed. The paper is structured as follows. We first give an overview of the two experimental data sets that our modeling studies are based on. Then the relationship between frequency and processing difficult is discussed and a parsing model based on probabilistic context-free grammars is proposed. We train and test the model on a standard corpus and demonstrate that it achieves broad coverage and good parsing accuracy on an unseen test set. Then the model is evaluated by correlating its predictions with the two experimental data sets. The model makes predictions about the differential behavior of verb final and verb initial sentences and provides evidence for the importance of lexical information in sentence processing.

Experimental Data

Keller (2000a, 2000b) presents experimental results on word order variation in subordinate clauses in German. These data form the basis for our modeling studies, and will be summarized in the following section.

Experiment 1

Keller's (2000a) experiment included transitive verbs such as *kaufen* 'buy' which take an animate subject and an inanimate object. Four different word order patterns were investigated: SOV, OSV, VSO, VOS (we use 'S' for subject, 'O' for object, and 'V' for verb). The object and the subject were realized as either a full NP or as a pronoun. The target word order was presented as a subordinate clause embedded by verbs like *glauben* 'believe'. (1) illustrates the SOV order with two full NPs.

(1) Maria glaubt, dass der Vater den Wagen kauft.
Maria believes that the[nom] father the[acc] car buys
'Maria believes that the father is buying the car.'

In the syntactic literature on German (e.g., Müller, 1999) SOV is generally regarded as the basic word order for subordinate clauses. Verb initial orders are regarded as ungrammatical, while scrambling (the permutation of subject and object) is regarded as marked, i.e., of reduced grammaticality, but it not outright ungrammatical.

Keller (2000a) used magnitude estimation (ME; Bard, Robertson, & Sorace, 1996) to test these theoretical claims. He elicited acceptability judgments for 16 word orders, each of which as represented by eight lexicalizations, yielding a total set of 128 sentences. Twenty native speakers of German were used as subjects. Under the assumption that ME judgments provide an index of global processing difficulty, Keller's (2000a) results support the following generalizations:

(2) a. Verb initial orders are harder to process than verb final orders.
b. Object initial orders are harder to process than subject initial orders.
c. Orders in which non-pronouns precede pronouns are harder to process than orders in which pronouns precede non-pronouns.

There is a range of experimental results that were obtained using different paradigms that confirm these processing preferences. This includes eye-movement and self-paced reading data (Bader & Meng, 1999; Scheepers, 1997), and a range of other comprehension and production paradigms (Pechmann, Uszkoreit, Engelkamp, & Zerbst, 1994).

Experiment 2

Keller (2000b, Experiment 6) extends these results to ditransitive verbs such as *vorstellen* 'introduce' that can

take three animate arguments. Six word order patterns were tested: SIOV, SOIV, ISOV, IOSV, OSIV, OISV, either with three full NPs or with two full NPs and one pronominalized NP ('O' denotes the direct object, 'I' the indirect object). Again, the word order was presented as a subordinate clause. (3) gives an example for the order SIOV with three full NPs.

(3) Ich weiß, dass der Manager dem Projektleiter
I know that the[nom] manager the[dat] project leader
den Mitarbeiter vorstellt.
the[acc] staff member introduces
'I know that the manager is introducing the staff member to the project leader.'

ME judgments were elicited for 24 orders; each order was presented in eight lexicalizations, yielding an overall set of 192 stimuli. Twenty-five native speakers were used as subjects. The results support claims (2b) and (2c), and yield the additional finding in (4) (verb initial orders were not included, hence the result in (2a) could not be tested).

(4) Orders in which the direct object precedes the indirect object are harder to process than orders in which the indirect object precedes the direct object.

We will use the data sets from Experiments 1 and 2 to test a model of global processing difficulty. The underlying assumption is that the ME score of a sentence can serve as a measure of processing difficulty: the harder a sentence is to process, the more unacceptable it is in an ME judgment task. Evidence for this assumption is provided by Bard, Frenck-Mestre, Kelly, Killborn, and Sorace's (1999) study which shows that ME data are correlated with data from self-paced reading and eye-tracking experiments.

Frequency and Processing Difficulty

The aim of the present study is to test the Tuning Hypothesis, i.e., the claim that there is a correlation between the frequency of a structure in the linguistic environment and global processing difficulty, as measured by magnitude estimation studies. Before we can test this claim, we have to find a way of approximating the linguistic environment of a speaker, i.e., we need a sample of the linguistic input that the speaker is exposed to. Such samples are readily available for a number of languages in the form of *corpora*, large computerized collections of text or speech. In the remainder of the paper, we will assume that the frequency distributions in the linguistic environment can be approximated by frequency distributions in a corpus.

For German, a suitable corpus is available in the form of Negra (Skut, Krenn, Brants, & Uszkoreit, 1997), a 350,000 word corpus of newspaper text. Negra is annotated with (a) part of speech labels (e.g., KOUS for complementizer, ART for article, NN for count noun, and VVFIN for finite verb); (b) syntactic information in the form of syntactic trees and phrase structure labels (e.g., NP for noun phrase, S for sentence); (c) grammatical function labels (e.g., HD for head, SB for subject, OA for accusative object). Figure 1 shows a Negra-style structure for the sentence in (1) (subordinate clause only; a subset of the Negra function labels is displayed).

A corpus annotated with syntactic structure affords a straightforward way of testing the Tuning Hypothesis:

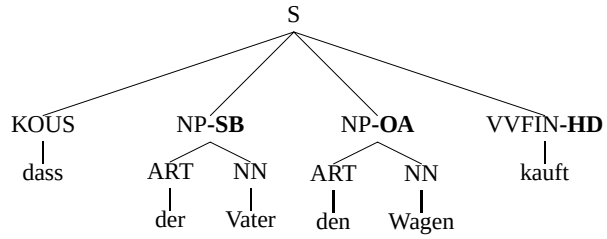

Figure 1: Example of a Negra tree

Verb	Total	IO order	OI order
geben 'give'	560	16	0
vorstellen 'present'	38	0	3
zur Verfügung stellen 'make available'	24	0	3

Table 1: Word order frequencies for full NPs in Negra reported by Kurz (2000)

We extract all instances of a given structure from the corpus and then correlate the resulting frequencies with a measure of processing difficulty.

This naive approach, however, is not feasible due to *data sparseness*: many of the word orders tested in Experiments 1 and 2 are rare in the corpus, which makes it difficult to obtain reliable frequency counts for these structures. As an example consider the frequency data that Kurz (2000) extracted from Negra. Kurz (2000) investigated the word order patterns of certain ditransitive verbs, such as *geben* 'give' *vorstellen* 'present', and *zur Verfügung stellen* 'make available'. All instances of these verbs were extracted from Negra and then classified with respect to the order of their objects. The resulting frequencies are given in Table 1. The column headed Total lists the overall frequency of a given verb in the corpus; the IO column contains the number of cases where the indirect object (dative NP) precedes the direct object (accusative NP), the OI column lists the counts for the inverse order. The frequencies in these two columns only take into account cases in which both objects are realized as full NPs; Kurz (2000) does not report data on pronominalized word orders.

In principle, we could now correlate the frequencies in Table 1 with the results of Experiment 2, in which it was found that OI orders are harder to process than IO orders (see (4)). However, as the corpus counts are too sparse for such a comparison: three out of six counts are zero, and also the other three orders are very rare. This means that no reliable claims can be derived from these counts.

This example indicates that data sparseness makes it impossible to directly correlate corpus frequency and processing difficulty, at least for the particular syntactic construction we are interested in here, and for the corpus we are using. In cognitive terms, this means that it is implausible to assume that the human parser directly keeps track of structural frequencies in its environment to determine word order preferences; it simply does not encounter the relevant structures often enough to derive reliable statistics.

S	→	KOUS NP NP VVFIN	.9
S	→	KOUS NP VVFIN	.1
NP	→	ART NN	1.0
VVFIN	→	kauft	1.0
KOUS	→	dass	1.0
ART	→	der	.8
ART	→	den	.2
NN	→	Vater	.6
NN	→	Wagen	.4

Figure 2: Example of a PCFG

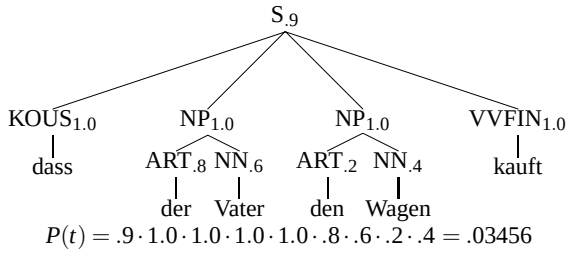

Figure 3: Example of tree generated by a PCFG

Probabilistic Context-Free Grammars

In the last section, we argued that data sparseness makes it implausible that the human parser directly records structural frequencies. We will therefore pursue an alternative hypothesis: that the parser keeps track of the frequencies of grammar rules.

More specifically, we will assume that global processing difficulty can be modeled using a probabilistic context-free grammar (PCFG). Models of syntactic disambiguation (i.e., of local processing difficulty) based on PCFGs have been proposed by a variety of authors and have been shown to account for experimental findings on human disambiguation preferences (Jurafsky, 1996; Crocker & Brants, 2000; Hale, 2001). We will extend this approach to global processing difficulty as measured by magnitude estimation studies.

A PCFG consists of a set of context-free rules, where each rule $LHS \rightarrow RHS$ is annotated with a probability $P(RHS|LHS)$. This probability represents the likelihood of expanding the category *LHS* to the categories *RHS*. In order to obtain a mathematically sound model, the probabilities for all rules with the same lefthand side have to sum to one. The probability of a parse tree T is defined as the product of the probabilities of all rules applied in generating T. We will assume that this probability is correlated with processing difficulty, i.e., that improbable structures are harder to process than probable ones.

An example for a PCFG is given in Figure 2. This grammar contains all the rules required to generate (1). Figure 3 displays the parse tree for this sentence, annotated with rule probabilities. The overall probability of the parse is also listed; it is computed as the product of all the rule probabilities.

A simple PCFG such as the one in Figure 2 has a number of obvious limitations; the linguistics distinctions it makes are not fine-grained enough. For instance, all noun phrases are assigned the category NP, even though the relative order of subjects, direct objects, and indirect objects was shown to trigger differences in processing behavior in Experiments 1 and 2. This problem can be addressed by incorporating *grammatical function information* into the grammar. For example, the noun phrase label NP can be split up into the labels NP-SB, NP-OA, and NP-DA for subjects, direct objects, and indirect objects, respectively. Grammatical functions are marked up in the Negra corpus, as illustrated in Figure 1. We can use this information to replace a PCFG rule such as S → KOUS NP NP VVFIN with a set of rules incorporating grammatical function labels, e.g., S → KOUS NP-SB NP-OA VVFIN-HD, and S → KOUS NP-OA NP-SB VVFIN-HD. The resulting grammar is more adequate for modeling processing differences that arise from differences in the order of subjects and objects (as they were demonstrated by Experiments 1 and 2).

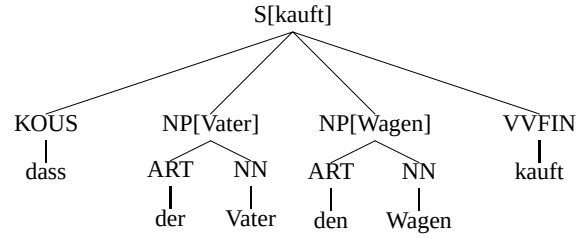

Figure 4: Example of tree generated by a lexicalized PCFG

An alternative way of improving the adequacy of a PCFG is adding *lexical information*. This approach has been pursued extensively in the computational linguistics literature (e.g., Carroll & Rooth, 1998) and has been shown to dramatically improve parsing performance. Lexicalization means that the set of category labels is extended by incorporating information about the head of the categories. For example, the category NP is split up into NP[Vater] and NP[Wagen] for noun phrases headed by the lexical items *Vater* 'father' and *Wagen* 'car', respectively. This is illustrated in Figure 4, which contains the lexicalized version of the tree in Figure 3. As with grammatical functions, this leads to an extension of the set of rules, yielding rules such as S[kauft] → KOUS NP[Vater] NP[Wagen] VVFIN.

Lexicalization can be seen as a way of approximating linguistic information that is not explicit in the corpus (and in the grammar). An example is *morphological information*: the label NP[Vater] implicitly contains the information that the NP is nominative, third person, and singular, i.e., the morphological features of *Vater* 'father'. Lexicalization is also a way of incorporating *co-occurrence information* into the grammar. The rule S[kauft] → KOUS NP[Vater] NP[Wagen] VVFIN contains the information that the verb *kauft* 'buys' co-occurs with the nouns *Vater* 'father' and *Wagen* 'car', thus capturing the semantic relationship between these three words. There is evidence in the psycholinguistic literature showing that the human parser makes use of morphological information (e.g., Trueswell, 1996) and semantic plausibility (e.g., Garnsey, Pearlmutter, Myers, & Lotocky, 1997).

Modeling Broad Coverage

We developed three models of global processing difficulty, as outlined in the previous section: (a) a *baseline model* using a standard PCFG, the (b) a *lexicalized model*

using a lexicalized PCFG, and (c) a *functional model* based on a PCFG that includes grammatical function labels. In all three cases both the grammar rules and the probabilities were derived from the Negra corpus.

Method

Negra was split into three subsets: the first 90% of the corpus were used as training set, the remainder was divided into a 5% test set and a 5% development set (for parameter tuning). Sentences with more than 40 words were removed (to increase parsing efficiency).

The baseline model was realized using the probabilistic left-corner parser Lopar (Schmid, 2000), running in unlexicalized mode. A grammar and a lexicon were read off the Negra training set, after empty categories and function labels had been removed from the corpus. The lexicalized model was also realized using Lopar, which in lexicalized mode implements Carroll and Rooth's (1998) model. Lexicalization requires that each rule in a grammar has one of the categories on its righthand side annotated as the head. For the categories S, VP, AP, and AVP, the head is marked in Negra. For the other categories, we used rules to heuristically determine the head. The functional model was implemented by inducing a new grammar from the training set: the function labels SB, OA, and DA were kept, but all other function labels were removed. The model was again implemented by running Lopar in unlexicalized mode.

The parameters for all three models were estimated using maximum likelihood estimation. This means that $P(LHS \rightarrow RHS)$, the probability of rule $LHS \rightarrow RHS$, is estimated as $P(LHS \rightarrow RHS) = f(LHS \rightarrow RHS)/N$, where $f(LHS \rightarrow RHS)$ is the number of times the rule occurs in the training data, and N is the overall number of rules in the training data. Various smoothing schemes are implemented in Lopar to address data sparseness, see Schmid (2000) for details.

Results

All models were evaluated by running them on the test corpus, which had remained unseen during model development. As is standard in the computational linguistics literature, we measured labeled bracketing: to score a hit, the parser has to predict both the bracket (the beginning or end of a phrase) and the category label correctly. We report labeled recall (LR), i.e., the number of correct labeled brackets found by the parser divided by the total number of labeled brackets in the test corpus, and labeled precision (LP), i.e., the number of correct labeled brackets found by the parser divided by the total number of labeled brackets found by the parser. We also list the F-score, which is defined as $F = 2 \cdot LP \cdot LR/(LP + LR)$.

The results are given in Table 2. The baseline model achieves an F-score of 71.9%, while the functional model performs slightly worse with an F-score of 70.9%. The lexicalized model performs worse than the baseline with an F-score of 63.9%.[1]

[1]For a detailed analysis of why standard lexicalized parsing models perform do not perform well for German, see Dubey and Keller (2003).

	LR	LP	F-score
Baseline	72.5	71.3	71.9
Lexicalized	67.2	60.9	63.9
Functional	71.9	70.0	70.9

Table 2: Testing the coverage of the models using labeled bracketing scores

Failed parses:	constant			removed		
	r	p	N	r	p	N
Baseline	.393	.000	128	.489	.000	109
Lexicalized	.512	.000	128	.636	.000	109
Functional	.137	.123	128	.374	.002	68

Table 3: Modeling results for Experiment 1

Modeling Global Processing Difficulty

Recall that the aim of this study is to test the hypothesis that there is a correlation between the frequency of a structure in the linguistic environment and global processing difficulty. We have argued that structural frequency cannot be measured directly due to data sparseness. Instead, we hypothesized that the probability of a structure, as computed by a PCFG, is a predictor of processing difficulty. In this section, this hypothesis will be tested against the data sets from Experiments 1 and 2.

Experiment 1

We tested each of the models (baseline, lexicalized, and functional) against the experimental data as follows. The sentences used as experimental materials was parsed by the model, and the probability of the most probable parse was computed. This probability was normalized by sentence length (measured as the number of words in the sentence). This is necessary as a PCFG assigns lower probabilities to longer sentences (all other factors being equal), as longer sentences involve more rule applications.

To compare the models, we conducted a set of correlation analyses: we correlated the log of the probability predicted by the model for each sentence with the log of the mean magnitude estimation score for this sentence. The parser failed on some sentences, i.e., it did not find a parse. There are two ways of dealing with this problem: (a) setting the probability of a failed parse to a small, constant probability, and (b) removing the failed parses from the data. Table 3 lists the correlation coefficients for all three models and for both ways of dealing with failed parses.

The results show that the baseline and lexicalized models obtain significant correlations with the ME data. The functional model only achieves a correlation once the failed parses are removed; however, there are a lot of failed parses here (only 78 data points remain), so this result has to be interpreted with caution. We performed a t-test to compare the correlation coefficients achieved by the baseline model and the lexicalized model (failed parses removed). The correlation of the lexicalized model was significantly higher ($t(109) = 2.454$, $p < .05$). The fact that the lexicalized model outperforms the unlexicalized baseline model points to the important role that morphological and semantic information (plausibility) plays for global processing difficulty. Such information is approximated in the lexicalized model, but

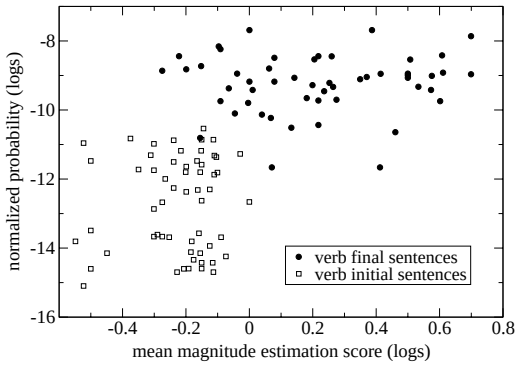

Figure 5: Correlation for the lexicalized model (failed parses removed) for Experiment 1

Failed parses:	constant			removed		
	r	*p*	*N*	*r*	*p*	*N*
verb final						
Baseline	.219	.082	64	.057	.690	52
Lexicalized	.233	.064	64	.067	.637	52
Functional	.155	.220	64	.296	.094	33
verb initial						
Baseline	.109	.392	64	.070	.605	57
Lexicalized	.123	.333	64	.089	.510	57
Functional	.044	.733	64	.011	.951	35

Table 4: Modeling results for Experiment 1, separation of verb final and verb initial items

not in the baseline model.

An inspection of the data shows the following interesting pattern. The models generate plausible analyses for verb final sentences and assign them high probabilities. No plausible analyses are found for verb initial sentences, and they are assigned low probabilities. This finding makes an interesting prediction with respect to processing difficulty as recorded by the ME judgments task: verb initial orders are predicted to receive low ME scores (recall that they are generally considered ungrammatical in the theoretical literature), while verb final orders should be assigned high ME scores. This prediction is borne out, as illustrated in Figure 5, which plots sentence probabilities against ME scores for the lexicalized model (with failed parses removed). We observe a clear separation of verb final and verb initial sentences. (The plots for the other two models show a similar pattern.)

The distinction between verb final and verb initial orders is captured successfully by our model. However, this raises the question if the model is able to predict processing difficulty also if *only* verb final or verb initial orders are tested. We investigated this by performing separate correlation analyses for the two subsets of the data. The results are given in Table 4. For the verb final sentences, we failed to find any significant correlations between probabilities and ME scores. The only exceptions were the baseline and the lexicalized models (failed parses set to a constant). Here a marginal correlation is obtained. (The correlation for the functional model with failed parses removed was also marginal, but included only 33 data points.) For verb initial sentences, the correlation coefficients were even lower and non-significant across the board.

Failed parses:	constant			removed		
	r	*p*	*N*	*r*	*p*	*N*
Baseline	.154	.033	192	.167	.022	190
Lexicalized	.208	.004	192	.230	.001	190
Functional	.441	.000	192	.019	.910	39

Table 5: Modeling results for Experiment 2

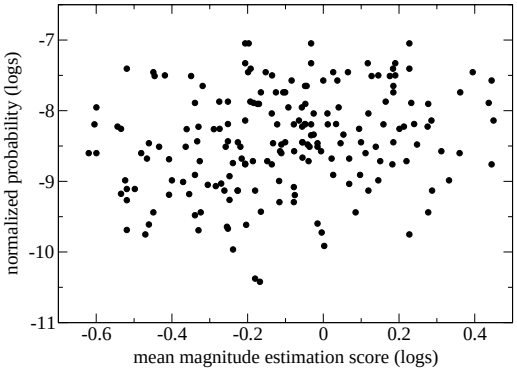

Figure 6: Correlation for the lexicalized model (failed parses removed) for Experiment 2

Experiment 2

The modeling results for the data from Experiment 1 showed that our models are able to distinguish between verb final and verb initial orders. However, the results also indicated that the models failed to reliably predict processing difficulty if the two word orders are considered separately. In the following, we will further investigate the behavior of the models for verb final orders based on the data from Experiment 2 (which only dealt with verb final sentences). Experiment 2 also provides a larger amount of data, viz., 192 items (compared to the 64 verb final items from Experiment 1). Furthermore, Experiment 2 provides a test of the generality of the modeling results, as it included ditransitive verbs, giving rise to additional word orders not included in Experiment 1.

We used the same modeling procedure as for the data from Experiment 1. The results for all three models and for the two ways of treating failed parses (set to a constant or remove) are listed in Table 5. We observe significant correlations for the baseline model and the lexicalized model for both ways of treating failed parses. Note that the lexicalized model again achieves higher correlation coefficients than the baseline model. However, a *t*-test shows that this difference is not significant.

For the functional model, we find a high correlation if failed parses are set to a constant; however, this correlation disappears if the failed parses are removed from the data set, leaving only 39 items. In other words, the functional model fails to assign a parse in most of the cases; this is probably due to data sparseness: there are not enough instances of ditransitive verbs in the corpus to acquire realistic probabilities for the functional model.

Figure 6 plots the magnitude estimation scores of Experiment 2 against the probabilities predicted by the lexicalized model (failed parses removed). The graph shows that this time all sentences behave in the same way; there is no split into two sentence types (such as the verb final/verb initial split in the model of Experiment 1).

Conclusions

In this paper, we tested the Tuning Hypothesis by applying it to global processing difficulty as measured by magnitude estimation. Two experimental data sets were used, both dealing with word order variation in German. We argued that data sparseness makes it implausible that the human parser directly records structural frequencies; instead, we assumed that it keeps track of rule frequencies. We implemented this hypothesis in a series of models based on probabilistic context free grammars. Sentence probabilities predicted by these models were shown to be significantly correlated with the ME scores obtained experimentally. We also showed that our models make predictions with respect to the differential processing of verb final and verb initial sentences. These predictions were borne out in the ME data.

Three different probabilistic models were tested, each incorporating different types of linguistic information. The baseline model was based on a standard PCFG; it achieved a significant correlation with ME data for both experiments. However, a model that incorporates lexical information into the category labels achieved a better fit with the experimental data than the baseline model. This indicates that the information that is approximated by the lexical model, viz., morphological information and semantic plausibility, plays a role in determining global processing difficulty. We also investigated the behavior of a functional model, in which the categories of the grammar incorporate grammatical function labels (subject, object, etc.). However, this model did not outperform the baseline model. This could be due to sparse data: there are not enough corpus examples available to reliably estimate the parameters of the functional model.

Many models in the psycholinguistic literature are designed to account for 'pathological' behavior of the processor, i.e., for cases of processing breakdown (such as garden paths). They are unable to explain why the human parser is generally very efficient and accurate on naturally occurring text, as pointed out by Crocker and Brants (2000). The models we presented here, however, are broad coverage models. This means they are able to parse naturally occurring, previously unseen text and achieve good parsing performance (precision and recall). They therefore capture this important characteristic of the human parser.

Finally, our approach also presents a methodological innovation. Previous work on probabilistic models of human parsing (Jurafsky, 1996; Crocker & Brants, 2000; Hale, 2001) only contained a simple, qualitative form of evaluation. Typically, the model is applied on a small set of examples to check if it predicts the correct processing pattern for these sentences. The present paper presented a more rigorous form of evaluation: we used two large data sets (128 and 192 sentences, respectively) and performed a quantitative evaluation by correlating model probabilities and ME scores. Model fit was measured using correlation coefficients, making it possible to compare the performance of different models on the same data.

References

Bader, M., & Meng, M. (1999). Subject-object ambiguities in German embedded clauses: An across-the-board comparison. *Journal of Psycholinguistic Research*, *28*(2), 121–143.

Bard, E. G., Frenck-Mestre, C., Kelly, L., Killborn, K., & Sorace, A. (1999). *Judgement and perception of gradable linguistic anomaly.* (Unpubl. ms., Human Communication Research Centre, University of Edinburgh)

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*(1), 32–68.

Brysbaert, M., & Mitchell, D. C. (1996). Modifier attachment in sentence parsing: Evidence from Dutch. *Quaterly Journal of Experimental Psychology*, *49A*(3), 664–695.

Carroll, G., & Rooth, M. (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 36–45). Granada.

Crocker, M. (1999). Mechanisms for sentence processing. In S. Garrod & M. Pickering (Eds.), *Language processing.* Psychology Press, London.

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29*(6), 647–669.

Dubey, A., & Keller, F. (2003). Probabilistic parsing for German using sister-head dependencies. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.* Sapporo.

Garnsey, S. M., Pearlmutter, N. J., Myers, E. M., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93.

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.* Pittsburgh, PA.

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

Keller, F. (2000a). Evaluating competition-based models of word order. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 747–752). Mahwah, NJ: Lawrence Erlbaum Associates.

Keller, F. (2000b). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality.* Unpublished doctoral dissertation, University of Edinburgh.

Kurz, D. (2000). A statistical account on word order variation in German. In A. Abeillé, T. Brants, & H. Uszkoreit (Eds.), *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora.* Luxembourg.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1996). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research*, *24*(6), 469–488.

Müller, G. (1999). Optimality, markedness, and word order in German. *Linguistics*, *37*(5), 777–818.

Pechmann, T., Uszkoreit, H., Engelkamp, J., & Zerbst, D. (1994). *Word order in the German middle field: Linguistic theory and psycholinguistic evidence* (CLAUS Report No. 43). Department of Computational Linguistics, Saarland University.

Scheepers, C. (1997). *Menschliche Satzverarbeitung: Syntaktische und thematische Aspekte der Wortstellung im Deutschen.* Unpublished doctoral dissertation, University of Freiburg.

Schmid, H. (2000). *LoPar: Design and implementation.* (Unpubl. ms., Institute for Computational Linguistics, University of Stuttgart)

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing.* Washington, DC.

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*(4), 566–585.