Modeling Attachment Decisions with a Probabilistic Parser: The Case of Head Final Structures

Ulrike Baldewein (ulrike@coli.uni-sb.de)
Computational Psycholinguistics, Saarland University
D-66041 Saarbrücken, Germany

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

Abstract

We describe an incremental, two-stage probabilistic model of human parsing for German. The model is broad coverage, i.e., it assigns sentence structure to previously unseen text with high accuracy. It also makes incremental predictions of the attachment decisions for PP attachment ambiguities. We test the model against reading time data from the literature and find that it makes correct predictions for verb second sentences; however, the model is not able to account for reading times data for verb final structures because attachment preferences in our training data do not match those determined experimentally. We argue that this points to more general limitations with our type of probabilistic model when it comes to realizing processing strategies that are independent of the data the parsing model is trained on.

Introduction

Experimental results show that human sentence processing is sensitive to different types of frequency information, including verb frame frequencies (e.g., Garnsey et al. 1997), frequencies of morphological forms (e.g., Trueswell 1996), and structural frequencies (e.g., Brysbaert & Mitchell 1996). Probabilistic parsing models are an attractive way of accounting for this fact, as they provide a theoretically sound way of combining different sources of frequency information into a coherent model. Typically, these models are hybrid models, combining symbolic knowledge (e.g., phrase structure rules) with frequency information (e.g., rule probabilities gleaned from a corpus).

In particular, probabilistic parsers have been used successfully to model attachment decisions in human sentence processing. Early models demonstrated the viability of the probabilistic approach by focusing on a small selection of relevant syntactic constructions (Jurafsky 1996; Hale 2001). More recently, *broad coverage* models have been proposed (Crocker & Brants 2000; Sturt et al. 2003) that can deal with unrestricted text. These models are able to account for the ease with which humans understand the vast majority of sentences, while at at the same time making predictions for sentences that trigger processing difficulties.

However, existing probabilistic models deal exclusively with English data, and thus fail to address the challenges posed by the processing of head final constructions in languages such as Japanese (e.g., Kamide & Mitchell 1999) or German (e.g., Konieczny et al. 1997). In this paper, we address this problem by presenting a probabilistic model of human sentence processing in German. The model is broad coverage, i.e., it generates accurate syntactic analyses for unrestricted text. Furthermore, it makes predictions for PP attachment ambiguities for both head initial and head final sentences. The model consists of two probabilistic modules: a syntactic module that proposes an initial attachment, and a semantic module that evaluates the plausibility of the proposed attachment, and corrects it if necessary.

We evaluate our model on reading time data for PP attachment, i.e., for structures in which a prepositional phrase can be attached either to a noun phrase or a verb. In German, PP attachment ambiguities can occur in two syntactic configurations: in verb second sentences, the verb precedes the NP and the PP as it does in English (see (1)).

(1) Iris tröstete den Jungen mit dem Lied.
Iris comforted the boy with the song.
'Iris comforted the boy with the song.'

(2) (daß) Iris den Jungen mit dem Lied tröstete.
(that) Iris the boy with the song comforted.
'(that) Iris comforted the boy with the song.'

In verb final sentences (which occur as subordinate clauses), the NP and the PP precede the verb (see (2)). As sentence processing is incremental, this means that an attachment decision has to be made before parser reaches the verb (and the frequency information associated with it). These structures therefore provide an interesting challenge for probabilistic models of sentence processing.

Reading studies (e.g., Konieczny et al. 1997, whose materials we use) have shown that in verb second sentences, the PP is preferredly attached according to the subcategorization bias of the verb (as in English). In verb final sentences, where verb frame information cannot be accessed until the end of the sentence, the PP is preferentially attached to the NP site.

The Model

Our parsing model consists of two modules: one is a syntactic module based on a probabilistic parser, which also has access to a probabilistic verb frame lexicon. This module guarantees broad coverage of language data and a high accuracy in parsing unseen text. The other module is a semantic module that uses probabilistic information to estimate the plausibility of the analyses proposed by the syntactic module.

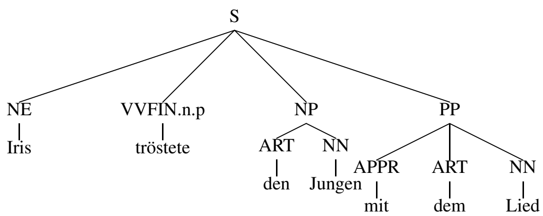
The model uses a syntax-first processing strategy: The syntactic module proposes a set of analyses for the input and ranks them by probability. The semantic module then computes the semantic plausibility of the analyses and ranks them by plausibility score. If there is a conflict between the decisions made by the two modules (i.e., the top-ranked analyses differ), this is interpreted as a conflict between syntactic preference and semantic plausibility and increased processing effort is predicted.

Syntactic Module

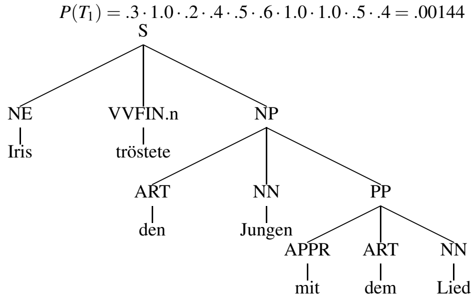
Modeling Syntactic Preferences The syntactic module consists of a probabilistic left-corner parser which relies on a probabilistic context free grammar (PCFG) as its backbone. A PCFG consists of a set of context-free rules, where each rule $LHS \rightarrow RHS$ is annotated with a probability $P(RHS|LHS)$. 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 left hand 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.

S	→	NE VVFIN.n.p NP PP	.3
S	→	NE VVFIN.n NP	.7
NP	→	ART NN	.4
NP	→	ART NN PP	.6
PP	→	APPR ART NP	1.0
VVFIN.n	→	tröstete	.8
VVFIN.n.p	→	tröstete	.2
ART	→	den	.5
ART	→	dem	.5
NE	→	Iris	1.0
NN	→	Jungen	.6
NN	→	Lied	.4
APPR	→	mit	1.0

Figure 1: Example of a PCFG



$P(T_1) = .3 \cdot 1.0 \cdot .2 \cdot .4 \cdot .5 \cdot .6 \cdot 1.0 \cdot 1.0 \cdot .5 \cdot .4 = .00144$



$P(T_2) = .7 \cdot 1.0 \cdot .8 \cdot .6 \cdot .5 \cdot .6 \cdot 1.0 \cdot 1.0 \cdot .5 \cdot .4 = .02016$

Figure 2: Example of trees generated by a PCFG

An example for a PCFG is given in Figure 1. This grammar contains the rules required to generate the two readings of (1). The readings are displayed in Figure 2, which also lists the parse probabilities, obtained by multiplying the probabilities of the rules used to generated a given tree.

This examples illustrates how PCFGs can be used for disambiguation: the two readings involve different rules (and rule probabilities), and therefore differ in their overall probabilities. In this example, reading T_2 is predicted to be preferred over T_1. Note that the grammar in Figure 1 incorporates verb frame probabilities: *tröstete* 'consoled' can either be a VVFIN.n (finite verb with an NP complement) or VVFIN.n.p (finite verb with an NP and a PP complement). The probabilities attached to these lexical items correspond to the psycholinguistic notion of *verb frame bias*, i.e., the probability of the verb occurring with a given subcategorization frame. The overall probability of an analysis is determined not only by verb frame bias, but also by structural probabilities attached to the phrase structure rules. This is a way of modeling structural disambiguation preferences (in this example, there is a bias for attachment to the NP). A PCFG therefore provides a principled way of integrating lexical preferences and structural preference, as argued by Jurafsky (1996).

Training and Test Data A PCFG is typically trained on a syntactically annotated corpus. For German, a suitable corpus is available in the form of Negra (Skut et al. 1997), a 350,000 word corpus of newspaper text. The Negra annotation scheme assumes flat syntactic structures in order to account for free word order in German. For example, there is no VP node dominating the main verb. Instead, subject, objects and modifiers of the main verb are its sisters, and all are direct daughters of the S node (see Figure 2). This means that scrambling phenomena simply alter the sequence of sisters in the tree, and do not involve movement and traces.

We checked the PP attachment preferences in Negra and found that in 60% of all sentences containing a verb and an NP object followed by a PP, the PP is attached to the verb. The corpus therefore reflects a general attachment preference for verb attachment. Additionally, we found that the subcategorization preferences for the verbs in our materials were reversed with regard to the preferences obtained by Konieczny et al. (1997) in a sentence completion task: the verbs that had a bias towards the NP-PP frame in the corpus exhibited an NP frame bias in the completion study, and vice versa.

For all subsequent experiments, 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 (used during model development). Sentences with more than 40 words were removed from the data sets (to increase parsing efficiency).

The syntactic module was realized based on Lopar (Schmid 2000), a probabilistic parser using a left-corner parsing strategy. A grammar and a lexicon were read off the Negra training set, after empty categories and function labels had been removed from the trees. Then the parameters for the model were estimated using maximum likelihood estimation. This means that the probability of a rule $LHS \rightarrow RHS$ is estimated as $P(LHS \rightarrow RHS) = f(LHS \rightarrow RHS)/N$, which is the number of times the rule occurs in the training data over the total number of rules in the training data. Various smoothing schemes are implemented in Lopar to address data sparseness, see Schmid (2000) for details. We also complemented the Negra verb frame counts with frame probabilities from an existing subcategorization lexicon (Schulte im Walde 2002), as the Negra counts were sparse.

Semantic Module

Modeling Semantic Plausibility The semantic module determines whether an attachment decision proposed by the syntactic module is semantically plausible by deciding whether the PP is more likely to be semantically related to the preceding verb or to the preceding noun.

Our semantic model rests on the assumption that "semantic plausibility" or "semantic relatedness" can be approximated by probabilistic measures estimated from corpus frequencies. Previous work provided evidence for this assumption by demonstrating that co-occurrence frequencies obtained from various corpora (including the web) are reliability correlated with human plausibility judgments (Keller & Lapata 2003).

Training and Test Data Ideally, the same training data should be used for the syntactic and the semantic module; however, this was not possible, as the semantic module requires vastly more training data. We therefore used the web to estimate the parameters of the frequency based measures (see Keller & Lapata 2003 for a detailed evaluation of the reliability of web counts). For the selectional preference method, we used one year's worth of text from the ECI Frankfurter Rundschau corpus as training data. This unannotated corpus is the basis for the Negra corpus, but it is much larger (34 million words). The corpus was parsed using a parser very similar to the syntactic module. Tuples of verbs and head nouns of modifying PPs were then extracted according to the structures assigned by the parser.

The development and test set for the semantic module were taken from the set of 156 items from Experiments 1 and 2 of Konieczny et al. (1997). The development set consists of 68 randomly chosen sentences, the remaining 88 sentences are used as a test set. The items from Experiment 1 vary word order (verb second and verb final), verb subcategorization preference (bias for NP frame or for NP-PP frame), and attachment (to the NP or verb), which is disambiguated by the semantic implausibility of one alternative. In Experiment 2, verb subcategorization preference was not varied. The development set was used to compare the performance of different semantic and syntactic models and to set the parameters for one semantic models. The final performance will be reported on the unseen test set.[1]

Plausibility Measures In computational linguistics, a standard approach to PP attachment disambiguation is the use of configuration counts from corpora (e.g., Hindle & Rooth 1991). To decide the attachment of n_{PP}, the head noun of the PP, to one of the attachment sites (the verb v or n_{NP}, the noun phrase), we compare how probable each attachment is based on previously seen configurations involving n_{PP} and the attachment sites. In many approaches, the the preposition p is also taken into account.

As outlined above, we used web counts to mitigate the data sparseness that such a model is faced with. In this approach, corpus queries are replaced by queries to a search engine, based on the assumption that the number of hits that the search engine returns is an approximation of the web frequency of the word in question (Keller & Lapata 2003). Of course text on the web is not parsed, which makes it difficult to identify the correct syntactic configurations. We follow Volk (2001) in assuming that string adjacency approximates syntactic attachment reasonably well, and simply use queries of the form `"V PP"` and `"NP PP"`. The search engines used were `www.altavista.de` and `www.google.com` (restricted to German data). Google generally outperformed AltaVista (presumably because it indexes more pages); the results reported below were obtained using Google counts.

We experimented with a variety of plausibility measures (*site* ranges over the two attachment sites, v and n_{NP}):

(a) $\frac{f(site,p)}{f(site)}$, the Lexical Association Score (LA), computes how likely the attachment site is to be modified by a PP with the preposition p.

(b) $f(site, p, n_{\mathrm{PP}})$, Model 1 of Volk (2001), relies on the raw trigram co-occurrence frequencies to decide attachment.

(c) $\frac{f(site,p,n_{\mathrm{PP}})}{f(site)}$, Model 2 of Volk (2001), takes into account that high-frequency attachment sites are more likely to co-occur with PPs.

(d) $log_2\left(\frac{f(site,n_{\mathrm{PP}})}{f(site)f(n_{\mathrm{PP}})}\right)$, Pointwise Mutual Information (MI) measures how much information about one of the items is gained when the other is seen. This measure has previously been used for the related problem of identifying collocations (words that appear together more often than chance, Church & Hanks 1990).

(e) $\frac{f(site,n_{\mathrm{PP}})}{f(site)} \cdot \frac{f(site,n_{\mathrm{PP}})}{f(n_{\mathrm{PP}})}$, Combined Conditional Probabilities (CCP) is similar to MI. It squares the joint probability term to give it more weight.

As will be explained below, we experimented with these measures in isolation, but we also combined them with Clark & Weir's (2002) approach for computing selectional preference from corpora. This approach relies on a lexical data base to compute the semantic similarity between lexical items.

Results

Syntactic Module

As mentioned in the introduction, the present modeling effort was guided by the idea of building a broad coverage model, i.e., a model that explains why human sentence processing is effortless and highly accurate for the vast majority of sentences; at the same time, the model should account for psycholinguistically interesting phenomena such as processing difficulties arising from attachment ambiguities. Incrementality is crucial for predictions of this type. In its original form, the Lopar parser used for the syntactic module is not incremental and was therefore modified to achieve partial incrementality. It now outputs its ranking of the attachment alternatives in two stages: after processing the PP and at the end of the sentence. This provides a record of incremental changes in the attachment preferences of the model when processing the critical region for which Konieczny et al. (1997) report eye-movement data (the noun of the PP in Experiment 1 and the PP object in Experiment 2).

To evaluate the broad coverage of the model, we ran the syntactic module on our unseen Negra corpus test set. The model was able to assign an analysis to 98% of the sentences. As is standard in computational linguistics, we tested the accuracy of the model by measuring 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, the number of correctly labeled brackets found by the parser divided by the total number of labeled brackets in the test corpus, and labeled precision, the number of correctly labeled brackets found by the parser divided by the total number of labeled brackets found by the parser.

The model achieved a labeled recall of 66.65% and a labeled precision of 63.92%. It is similar to the baseline model of Dubey & Keller (2003), who report a maximum labeled recall and precision of 71.32% and 70.93%.

To further evaluate the syntactic model, we tested it on the test set generated from Experiments 1 and 2 of Konieczny et al. (1997). This allows us to determine whether the syntactic module is able to correctly resolve the PP attachment ambiguities even without access to any semantic information.

Table 1 shows the parser's decisions at the PP for verb final and verb second sentences. We report the number and the percentage of correct attachments per condition. In the verb final condition of Experiment 1, the parser always attached the PP to the verb. No verb frame information is available to guide the decision when the PP is processed, so the baseline is random guessing (50%). In verb second sentences, the parser can use the subcategorization preference of the verb, which leads to the correct attachment in 50% of all cases. The parser indeed reaches this baseline. In Experiment 2, the parser again always attaches the PP to the unseen verb in verb final sentences. In the verb second condition, there is a marked preference to attach according to verb bias, but only 42% of attachments are correct over both conditions.

[1] Since for all models except one no parameters were set on the development set, we had to maintain a fixed development-test split to ensure the test set remained truly unseen.

	Verb final	Verb second
Experiment 1		
NP frame, V bias	7 (100%)	2 (29%)
NP frame, NP bias	0	5 (83%)
NP-PP frame, V bias	5 (100%)	3 (60%)
NP-PP frame, NP bias	0	2 (33%)
% correct	50%	50%
Experiment 2		
NP frame, V bias	9 (100%)	1 (11%)
NP frame, NP bias	0	5 (56%)
% correct	50%	33%
Baseline	50%	50%

Table 1: Syntactic module: correct attachment decisions at the PP for the test set from Experiments 1 and 2

Measure	CCP	MI	LA	Volk 1	Volk 2
Development Set					
# correct	23	22	17	22	21
% correct	67.6%	64.7%	50%	64.7%	61.7%
Test Set					
# correct	21	23	–	23	27
% correct	50%	54.8%	–	54.8%	64.3%
Baseline	50%	50%	50%	50%	50%

Table 2: Semantic module, verb second: results of the plausibility measures on the development and test set

Semantic Module

Verb Second Sentences As a next step, we evaluated the semantic module, again on the data derived from Experiments 1 and 2 of Konieczny et al. (1997). We again used the chance baseline (50%) that the syntactic module was unable to outperform.

The verb second sentences arguably constitute the standard case for PP attachment: Both possible attachment sites have been seen before the attachment has to be decided. In a first attempt, the five plausibility measures introduced above were tested on the development set. Table 2 shows that the CCP measure performed best, while the Lexical Association measure failed to beat the baseline. The CCP measure should therefore be chosen to model semantic attachment in verb second sentences. However, on the test set (see Table 2), the best and worst measure changed places. This time, the Volk 2 measure performed best. No measure significantly outperformed the others or the baseline.

As the performance of the CCP measure on the test set was disappointing, we experimented with a second approach that combines the Volk 2 model of PP attachment with a model of selectional restrictions. We used Clark & Weir's (2002) approach, which was extended to German by Brockmann & Lapata (2003), whose implementation we used. Relying on a semantic hierarchy (in our case: GermaNet, Hamp & Feldweg (1997)), the Clark & Weir algorithm finds the statistically optimal superclass (concept) for input nouns given a verb and the relation between noun and verb. The probability of a concept c given a verb v and relation rel is computed as:

(3) $P(c|v,rel) = P(v|c,rel)\frac{P(c|rel)}{P(v|rel)}$

To find the best concept for a $\langle n, v, rel\rangle$ triple, at each step up the hierarchy, the probability estimate for the new concept is compared to that of the original concept. When the estimates differ significantly, the lower concept is assumed for the noun. The parameters of this algorithm are the statistical test used (χ^2 or G^2) and the α value which determines the level of significance required for the test. The G^2 test proved more suitable for our task, while a variation of α value had no noticeable effect.

We used the development set to estimate a threshold value for the attachment decision. Coverage on the test set was only 48% due to sparse data. Whenever the Clark & Weir method did not return a value, we backed off to the decision made by the Volk 2 model (which is the most consistently performing model). Recall that this model has a 64% precision on the test data while the chance baseline is 50%. The combined model reaches 67% precision on the same data (precision for the selectional preference model alone is 70% for 48% of the data). This model performs best numerically (though not significantly so) and was used in the final model.

Verb Final Sentences A particularly interesting case arises with respect to verb final sentences (see (2)): at the critical region (once the PP has been processed), the verb is not available yet, which means that the plausibility of the combination of the verb with the head noun of the PP cannot be computed at this point. Konieczny et al. (1997) found processing difficulty in these cases when the PP was an implausible modifier of the noun, so apparently immediate semantic evaluation sets in and has to be accounted for.

In the verb final case, we therefore have to estimate the plausibility of the PP head noun modifying the NP as opposed to an unseen verb. One way of doing this is to average over the results for the PP head noun and every possible verb to obtain a generic value for verb attachment. We restrict ourselves to just the verbs in the test and development set. This backoff approach was realized for four models.

An alternative is to use the prior probability of the PP head noun as an estimate of its conditional probability with every possible verb. The prior probability of the PP head noun is its frequency divided by the size of the corpus, $f(n_{\mathrm{PP}}/N)$. In the case of web counts, N is the number of all documents searched. This figure was empirically estimated as proposed by Keller & Lapata (2003). Note that this method of backoff is possible only for the CCP measure, because the probabilities to be estimated for the other methods are too complex.

Table 3 gives the results on the development set for the items from Experiment 1. The items from Experiment 2 could not be tested as the averaging procedure is extremely costly in terms of web queries. The CCP model with simple backoff to the prior shows the best results at 60% correct attachments. We therefore used it for the final evaluation to predict attachments for verb final sentences. Table 3 also lists the results on the test set for items from Experiments 1 and 2. CCP with backoff to the prior performed better than most models that use averaging, and substantially outperforms CCP with averaging. The best model is Volk 2 with averaging. Again, no measure outperforms the baseline of 50% correct attachments or the other models.

	Prior	Average			
Measure	CCP	CCP	MI	Volk1	Volk2
Development Set					
# correct	12	11	8	9	11
% correct	60%	55%	40%	45%	55%
Test Set					
# correct	24	20	22	23	25
% correct	57.1%	47.6%	52.4%	54.8%	59.5%
Baseline	50%	50%	50%	50%	50%

Table 3: Semantic Module, verb final: Results on the development (Exp. 1) and test set (Exp. 1 and 2)

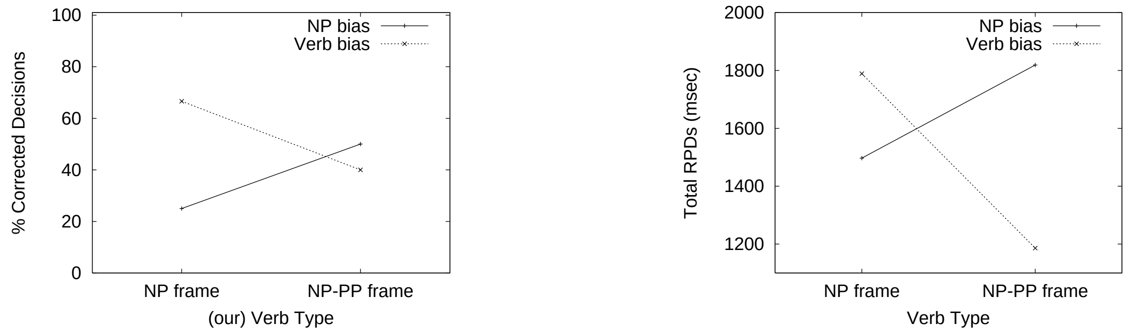


Figure 3: Exp. 1, verb second: Predictions of the combined model (left) compared to the Konieczny et al. (1997) data (right)

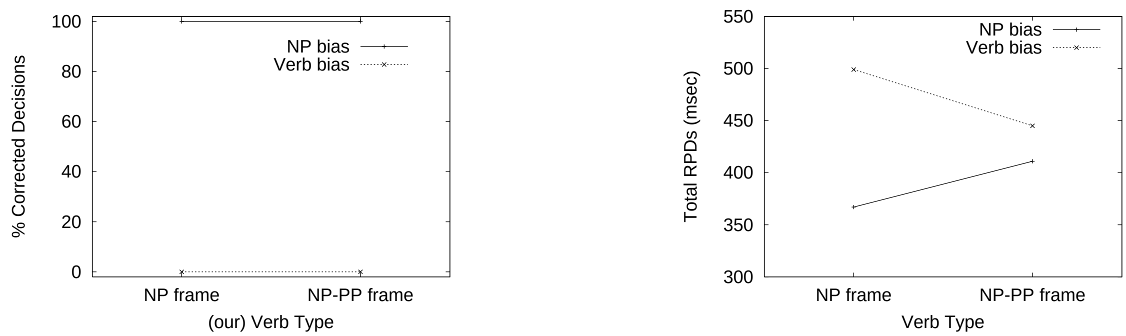


Figure 4: Exp. 1, verb final: Predictions of the CCP/Prior model (left) compared to the Konieczny et al. (1997) data (right)

Combined Model

In the previous sections, we evaluated the syntactic and semantic module separately. We found that the syntactic module performs at the level of the chance baseline of 50%, while the semantic module achieves an accuracy of up to 67% for verb initial sentences and 60% for the verb final sentences. A more interesting question is how well the model accounts for the processing difficulties that are evident in the eye-movement data reported by Konieczny et al. (1997). As mentioned at the beginning of the Results Section, our model makes predictions for the critical region used by Konieczny et al. (1997) (the PP). Recall also that we assume that a conflict between syntactic preference and semantic plausibility predicts increased processing effort.

As explained in the section on Training and Test Data above, the subcategorization variable was reversed for our data: where Konieczny et al. (1997) assume an NP-PP frame bias, we found a preference for the NP frame in our corpus (and vice versa). Below, our model's predictions are labeled with the preferences found in our data, while data from Konieczny et al. (1997) are labeled with the preferences they found. Figure 3 compares the predictions of our model with Konieczny et al.'s results in Experiment 1 for verb second sentences.[2] The graph for our model gives the percentage of correct decisions by the semantic module that are in conflict with the the decisions of the syntactic module. Such conflicts predict longer reading times, and the more conflicts in a condition, the higher we expect the average reading times to be. The figure shows that our model predicts the data pattern found by Konieczny et al. (1997) (who report regression path durations, RPDs).

[2]Note that our results are on the unseen subset of the items only, while the reading times are on all items.

In verb final sentences (Figure 4), the syntactic module always predicts verb attachment, so correct decisions for NP attachment by the semantic module always lead to a conflict. This pattern does not correspond to the Konieczny et al. (1997) reading data, which show a general preference to attach to the NP. Figure 5 shows a replication in principle of the reading time data in the verb second case. In Konieczny et al.'s (1997) pretests, all the verbs subcategorized for an NP and a PP, while in our data, they preferredly subcategorize for just an NP. Our model predicts longer reading times for the NP frame when subcategorization preference and semantic disambiguation are mismatched, which is what Konieczny et al.'s (1997) show for the NP-PP frame. The verb final case again fails: Instead of predicting preferred attachment to the NP (Matched bias for our data, Mismatched bias for Konieczny et al.'s data), the model predicts verb attachment.

Discussion

While our model replicates Konieczny et al.'s (1997) reading time results for PP attachment in the verb second case, it fails to account for reading times of verb final sentences. This failure is caused by the syntactic module which always predicts verb attachment in verb final sentences, while there is a human preference for NP attachment in these cases.

The behavior of the syntactic module is influenced by two factors. One is the probability of phrasal rules such as S $\rightarrow$ NE VVFIN.n.p NP PP. The second factor is a verb-specific frame bias, which manifests itself as probabilities for lexical rules such as VVFIN.n.p $\rightarrow$ *tröstete*. In verb second sentences, the verb's frame probability together with the phrasal rule probability determines the analysis proposed by the syntactic module. In verb final sentences, however, only the phrasal probabilities are used (as the verb is not yet avail-

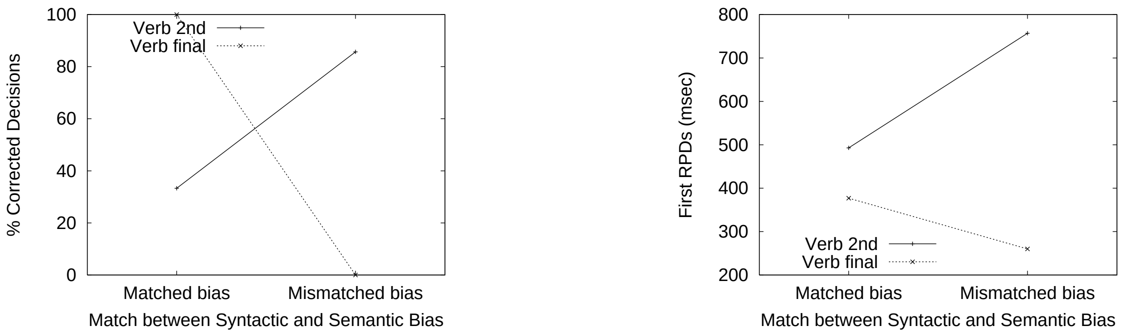


Figure 5: Exp. 2: Predictions of the combined model (left) compared to the Konieczny et al. (1997) data (right). Verbs subcategorize for an NP frame in our data and for an NP-PP frame in the Konieczny et al. data.

able), so the syntactic module makes the same prediction for all verb final sentences. This prediction is incorrect because the general PP attachment bias in the corpus is to the verb, rather than to the NP as in the reading time data.

This points to a more general problem with probabilistic models: They can only be as good as the training data. It is therefore vital to check relevant properties of the training corpus in comparison to experimental data when developing probabilistic models. Balanced corpora that consist of language data from different sources are more reliable in this respect than newspaper corpora such as the Negra corpus.

This means that the failure to model the verb final data is not a failure of probabilistic models per se; our approach would be in principle capable of modeling the general attachment preference to the NP in verb final sentences, if the attachment preference in the training data corresponded to that in the experimental results. Thus, our results strengthen the case for probabilistic models by showing that they can be applied even to head final constructions.

It is important to note, however, that our explanation of the German PP attachment data in terms of biases in the training corpus is at variance with explanations in the literature. For instance, Konieczny et al. (1997) proposes a strategy of *Parameterized Head Attachment* to explain why the parser prefers to attach incoming material (such as the PP) to existing sites (such as the verb). This strategy, which aims at the immediate semantic evaluation of the input, is designed to cope with head final structures in general, not only in the case of PP attachment. A basic PCFG model such as the one used in this paper is not able to implement such a general strategy.

Conclusions

We have presented a two-stage model parsing model that accounts for PP attachment in German. The model is able to assign correct sentence structures to unseen text and predicts average reading times in verb second sentences. For verb final sentences, the model fails to correctly predict the reading time data. The reason is that our training corpus exhibits a general bias for attaching PPs to the wrong attachment site (to the verb instead of the NP). In principle, however, our model would be able to account for the data in the verb final case if the training data were consistent with experimental findings. Our findings therefore strengthen the case for probabilistic models of language processing by showing their applicability to head final structures. At the same time, they demonstrate that probabilistic models can be highly sensitive to idiosyncrasies in the training data.

References

Brockmann, C., & Lapata, M. (2003). Evaluating and combining approaches to selectional preference acquisition. In *Proc. EACL*, (pp. 27–34), Budapest.

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

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22–29.

Clark, S., & Weir, D. (2002). Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, *28*, 187–206.

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

Dubey, A., & Keller, F. (2003). Probabilistic parsing for German using sister-head dependencies. In *Proc. ACL*, (pp. 96–103), 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. *J. of Memory and Language*, *37*, 58–93.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proc. NAACL*, Pittsburgh, PA.

Hamp, B., & Feldweg, H. (1997). GermaNet: A lexical-semantic net for German. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo, & Y. Wilks (eds.), *Proc. ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, (pp. 9–15), Madrid.

Hindle, D., & Rooth, M. (1991). Structural ambiguity and lexical relations. In *Proc. ACL*, (pp. 229–236), Berkeley, CA.

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

Kamide, Y., & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. *Language and Cognitive Processes*, *14*, 631–662.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, *29*, 459–484.

Konieczny, L., Hemforth, B., Scheepers, C., & Strube, G. (1997). The role of lexical heads in parsing: Evidence from German. *Language and Cognitive Processes*, *12*, 307–348.

Schmid, H. (2000). LoPar: Design and implementation. Unpubl. ms., IMS, University of Stuttgart.

Schulte im Walde, S. (2002). A subcategorisation lexicon for German verbs induced from a Lexicalised PCFG. In *Proc. LREC*, vol. IV, (pp. 1351–1357), Las Palmas, Gran Canaria.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proc. ANLP*, Washington, DC.

Sturt, P., Costa, F., Lombardo, V., & Frasconi, P. (2003). Learning first-pass structural attachment preferences with dynamic grammars and recursive neural nets. *Cognition*, *88*, 133–169.

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

Volk, M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proc. Corpus Linguistics*, Lancaster.