Modelling Attachment Decisions with a Statistical Parser

Ulrike Baldewein

Master of Science
Cognitive Science and Natural Language
School of Informatics
University of Edinburgh
2003

Abstract

Probabilistic models of human sentence processing have proved widely successful in modelling human attachment decisions. This Thesis describes a two-stage probabilistic model of human incremental parsing of German PP-Attachment. German verb final sentences provide a new opportunity for the study of PP-Attachment because the PP is processed in the absence of the sentence head. In this situation, it is preferentially attached to the (existing) NP site.

The model consists of two modules: A syntactic module based on a standard stochastic parser and a shallow semantic module that makes final attachment decisions on the basis of Web counts. Conflicts between decisions made by the modules are interpreted as predicting longer reading times. The model's predictions are compared to average reading times from an eyetracking study (Konieczny et al., 1997).

The model correctly accounts for attachment preferences in verb second sentences. This is a replication of results for English. It fails to account for preferences in verb final sentences. We argue that these preferences cannot be modelled by a probabilistic CFG at all. They probably belong to a range of phenomena that have been explained by either the initial influence of global structural frequencies or by the existence of general parsing strategies. Purely probabilistic CFGs cannot accomodate either of these explanations.

Acknowledgements

I am glad to have many people to thank for their help and support. First of all, I would like to thank someone who has been with me every minute of this year and who has given me endless care and guidance: Jesus, thank you for loving me first.

Those who helped with the Thesis In this section, thanks are due first of all to my supervisor Frank Keller. You have been incredibly helpful. It was a pleasure to work with you!

Helmut Schmid kindly allowed me access to the source code of LoPar, his left-corner parser.

Thanks also go to Andreas Eisele, who provided us with a list of morphological forms (generated with MMorph) for all words in the experimental items.

I would also like to thank Sabine Schulte im Walde for access to the subcategorisation lexicon.

Another thank you goes to Matt Smillie and Viktor Trón, who helped me untangle a few C-pointers, and to Markus Becker, with whom I have had many helpful discussions about LoPar.

Friends and Family Sebastian, thank you for all those hours on the phone and for taking me away on much-needed holidays! Thank you for being Just the Way you are...

Mama, you have been a great support! Thanks for just being a wonderful mother.

Beata and Sarah, thanks for being great friends! I am glad to have met you and I hope we will stay in touch.

There have been many others who have made this year very special – the fellow MScs in the lab, people at Buccleuch & Greyfriars Free Church, my friends at home who kept in touch. Thanks to all of them, too!

Funding Bodies I was funded by DAAD (Deutscher Akademischer Austauschdienst) through their one-year programmes in Great Britain.

I have also been supported by the Studienstiftung des deutschen Volkes.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Ulrike Baldewein*)

To the memory of Helmut Baldewein (1942 - 1995)

Table of Contents

List of Figures

Chapter 1

Introduction

The human sentence processing mechanism, its strategies and the factors that influence its workings are a focal point of research in the field of psycholinguistics. This research proceeds both by experiments and by computational models informed by the outcome of these experiments. While experimental data shed light on the performance of the sentence processor in a controlled setup, computational models allow us to verify the viability of general theories derived from a wide range of experimental data. They also allow an estimate of the relative strength and importance of factors that are known to influence the parsing process.

One such factor is the frequency of words and structures in readers' language experience. Frequency effects have been shown to be prevalent on all structural levels of sentence processing. For example, Duffy et al. (1988) demonstrate frequency effects in word sense disambiguation and Cuetos and Mitchell (1988) and Mitchell et al. (1995) find indications of the influence of frequency information on the phrase structure level.

The importance of frequency information in parsing is mirrored in the success of different kinds of frequency-based models of lexical disambiguation and parsing. These models are generally either constraint-based and inspired by research into neural networks (e.g. Spivey and Tanenhaus (1998); MacDonald et al. (1994)), or rule-based in the computational linguistics tradition (Jurafsky, 1996; Crocker and Brants, 2000; Sturt et al., 2003). Rule-based probabilistic models have the advantage of being able to draw on research in computational linguistics which has led to the construction of reliable, wide-coverage parsers that rely on probabilistic information (Collins, 1997;

Charniak, 2000). This allows them to explain not only pathological language phenomena which cause problems for the human processor, but also to account for the vast amount of language data that is being processed quickly and robustly.

This thesis further investigates the question to which extent the frequency of syntactic structures in readers' language experience predicts attachment decisions during the course of parsing. To answer this question, a standard computational linguistic parser is equipped with statistical information about the frequency of words and phrases in a training corpus. We then test how well the parser's initial attachment decisions correspond to the decisions humans evidently make in reading experiments.

More specifically, the task is to model human preferences in the attachment of prepositional phrases (PPs). Both noun phrases (NPs) and verbs can be freely modified by PPs, so in many cases there is a syntactic attachment ambiguity between attachment to an NP or the verb. On a syntactic level, the decision is influenced by whether the verb shows a preference for taking a PP object. This effect has been successfully modelled for English (Jurafsky, 1996; Crocker and Brants, 2000). We investigate the same effect for German. This gives us the opportunity to test the parser's prediction for the case where word order is as it is in English as well as looking at a case where the verb is the last word of the sentence and has therefore not been seen when the PP is processed. To our knowledge, this phenomenon has not yet received attention in psycholinguistic modelling.

Apart from syntactic influences, a second important factor in PP-attachment is of course the semantics of the attachment alternatives. The experimental data we attempt to model disambiguate the attachment by semantic plausibility. Therefore, a separate shallow semantic module was built whose task it is to make the final decision on attachment. Two approaches to the task are compared: One decides the attachment according to whether the noun in the PP has been more frequently attached to the NP in question or the verb in large amounts of language data. The other estimates semantic relatedness of the noun in the PP and the potential attachment sites from word co-occurrence in a large corpus and advocates attachment to the more related site.

The structure of the Thesis is as follows. Chapter 2 is an introduction to relevant previous work. It gives an overview over research into PP-Attachment in psycholinguistics, and goes on to present probabilistic models of human sentence processing. Frequency-based strategies of deciding PP-Attachment are also discussed because they are relevant for the development of the semantic disambiguation module. Chapter 3 summarises the task that is set for the model. Chapter 4 describes the development of the syntactic module, and Chapter 5 gives details of the semantic module and an outlook on future work.

The performance of the modules and of the model in general is evaluated and discussed in Chapter 6. Chapter 7 contains concluding remarks.

Chapter 2

Previous Work

2.1 PP-Attachment and Human Sentence Processing

Psycholinguistics is concerned with how the brain processes language. The human sentence processor works extremely quickly, accurately and robustly. Unfortunately, understanding language is an unconscious process, so there are only few ways of finding out how the processor goes about its work. One way to test its strategies is by analysing where it encounters difficulty and by inferring from those which processing principles the processor was following. Difficulty can be induced by either complexity or ambiguity of the input sentences, and measured by subjects' opinions about the acceptability of the sentences involved or by an increase in reading time for those sentences.

One basic finding is that sentence processing proceeds incrementally, that is word-by-word (as opposed to waiting for chunks of input to accumulate and then processing those). Further investigations have focused on which factors influence the incremental processing decisions and what the timecourse of these effects is. In this respect, PP-Attachment is a useful phenomon, because the final attachment decision is influenced by both lexical and semantic factors.

2.1.1 Prepositional Phrase Ambiguities

The attachment of prepositional phrase into sentences is a well-studied example of ambiguity. The ambiguity here arises from the fact that in a sentence like *The man saw the snake with the binoculars* the attachment of the prepositional phrase (PP) *with the binoculars* is syntactically permissible both to the noun phrase (NP) *the snake* and the verb *saw* (see Figure 2.1). The outcome of the attachment depends mainly on two factors, namely which configuration of objects (which *subcategorisation frame*) the verb prefers and which attachment is semantically more plausible. Assuming that *saw* prefers just a noun phrase as its object, the verb's subcategorisation preference would call for attachment of the PP into the NP. This corresponds to the lower attachment alternative in Figure 2.1. Since this attachment is semantically very implausible, the final structure sees the attachment of the PP to the verb.

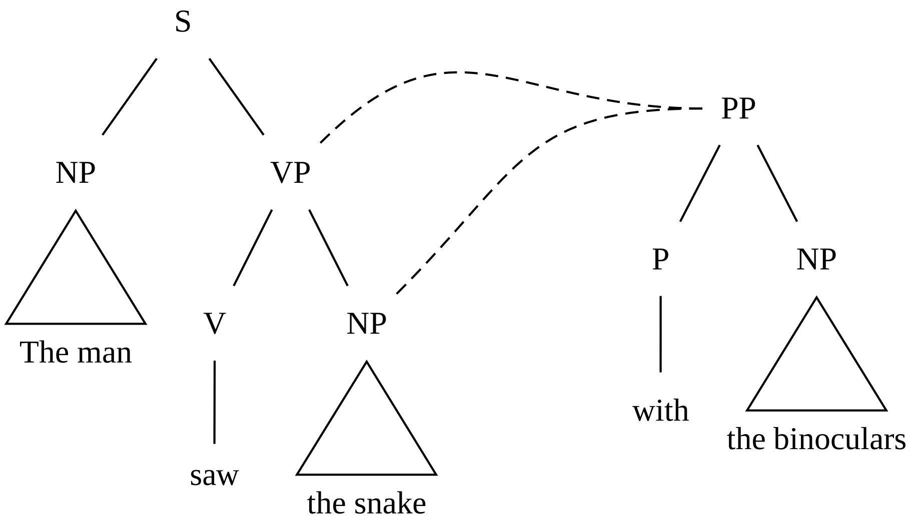

Figure 2.1: PP-Attachment Ambiguity

A verb's subcategorisation preferences seem to be accessed faster and weigh more strongly initially than semantic effects (Strube et al., 1989; Garnsey et al., 1997). Work by Schütze and Gibson (1999) generalises from constraints set by the verb's argument structure to a general preference to attach PPs as arguments rather than as adjuncts wherever there is a choice. Semantic effects include the plausibility of an attachment given world knowledge (e.g. it is very implausible to see *the snake with the binoculars*) or effects of disambiguation between several possible objects (Altmann and Steedman, 1988) as well as definiteness effects (Spivey-Knowlton and Sedivy, 1995). From a linguistic point of view, these latter effects are already of pragmatic nature. They appear when the experimental materials seemingly restate known facts or introduce previously unheard-of discourse entities as if they were known, thereby disobeying the Gricean discourse maxims of Quantity, “Make your contribution as informative as is required, do not make your contribution more informative than is required.” and Manner “Be perspicuous: Avoid ambiguity. Be brief.” (Grice, 1975).

Different languages appear to have different attachment preferences. For German, the default attachment is to the NP (Konieczny and Hemforth, 2000). For English, there is an online preference for VP-attachment (Frazier and Rayner, 1982).

2.1.2 Theories of Attachment

PP-Attachment is usually not so much investigated in its own right, but rather as a test case for more general hypotheses about the sentence processing mechanism. A very influential and much-contended theory of how human processing works is the Garden Path theory by Frazier and Rayner (1982). It rests on the two structural principles of *Minimal Attachment* and *Late Closure*. In brief, Minimal Attachment causes the parser to prefer structurally simpler analyses (that contain fewer nodes in the assumed grammar formalism) and Late Closure causes it to include new material into the phrase that was constructed last. If both principles are applicable, Minimal Attachment takes precedence.

Konieczny et al. (1997) advocate *Parametrised Head Attachment*, a theory of incremental parsing which states that a newly read head should preferentially be attached to an existing head such that a semantic interpretation of the sentence so far can be formed immediately. The first principle therefore is to attach a new head to any existing head. In the case of competition between several heads, the new item should be attached to the existing head that highlights an appropriate argument slot. If this does not fully disambiguate, attachment is made to the most recent head. The attachment preference to heads with an open argument slot emulates the effect of Frazier and Rayner (1982)’s Minimal Attachment in most cases, because the new item is attached high to the verb instead of to the object, which might be structurally more costly. The principle of attachment to the most recent head accounts for the recency effect e.g. in adverb attachment, because this case is not disambiguated by the potential attachment sites' frame preferences and is a re-statement of the Late Closure principle. Parametrised Head Attachment predicts that in the absence of a verb (e.g. in German verb final clauses), there should initially be a preference for NP-attachment, while verb subcategorisation frames decide the attachment behaviour for verb second sentences. Minimal Attachment, in contrast, would always predict an initial preference for verb attachment.

This theory is supported by results from Konieczny et al. (1995) and Konieczny et al. (1997). Konieczny et al. (1997), whose results I aim to replicate with my model, found an initial attachment preference for the NP in absence of the verb. Additionally, they found that when a verb is present that prefers an NP and a PP object, the PP is preferentially attached to that verb and a conflicting semantic bias leads to processing difficulty. If the verb prefers just an NP object, the PP is preferentially attached to the NP.

2.2 Probabilistic Models of the Human Sentence Processor

2.2.1 Probabilistic Context-Free Grammars

Most of the models introduced below rely on context-free grammars (CFGs) as a means of building structure over the words in the input sentences. A context-free grammar consists of a set of non-terminal symbols (phrase labels and word tags), a set of terminal symbols (the words) and rules of the form

$$Nonterminal \rightarrow (Nonterminal \mid Terminal)^{+}$$

That is, a non-terminal symbol on the left hand side can be rewritten as a string of non-terminal or terminal symbols on the right-hand side of the rule. For example, with the rule S $\rightarrow$ NP VP, a sentence can be rewritten as a combination of a noun phrase and a verb phrase. Remaining non-terminal symbols can in turn be rewritten by appropriate rules (e.g. NP $\rightarrow$ Det N). The rewriting process can be visualised in form of a tree, where every left-hand side is a node in the tree and the right-hand side symbols are daughters of this node. Since there is no notion of the context in which the left-hand side symbol appears, CFGs cannot immediately deal with phenomena such as long-distance dependencies, where displaced material has to be linked back to its mother node, e.g. via a trace.

A probabilistic CFG (PCFG) enhances this framework by assigning a probability to every grammar rule. Probabilities for phrases and sentences are compiled by multiplying the rule probabilities involved in building them. The probability assigned to completed structures allows an estimation of how probable one structure is in comparison to other possible structures. Here, it is assumed that highly probable structures are more acceptable than improbable ones. The probabilities are usually gained from structure counts in corpora. PCFGs can only assign meaningful probabilities to completed phrases. In order to arrive at a well-formed probability distribution, the probabilities of all rules with the same left-hand side have to sum to one and there cannot be unbounded recursion (i.e. one symbol being directly or indirectly rewritten as itself).

2.2.2 Models

As introduced in Section 2.1 above, important properties of the human sentence processor that should be modelled are

- Correct and effortless processing of most language data
- Incremental processing
- Difficulty for a relatively small set of ambiguity phenomena

The decision to use probabilistic models to model these three points is motivated by both empirical and theoretical considerations. Firstly, empirically, the human processor has been shown to be sensitive to frequency effects. For example, Duffy et al. (1988) have demonstrated frequency effects in lexical disambiguation. Trueswell (1996) shows a sensitivity to verbs' preferred tenses. On the higher syntactic level, the existence of different default attachments for modifiers crosslingually and even within languages[1] indicates that modifier attachment preferences are not the byproduct of some general processing strategy. The *Tuning Hypothesis* (Mitchell et al., 1995) proposes to explain these data by the parser's use of statistical information about preferred attachment configurations that influences later decisions.

Chater et al. (1998) present a theoretical investigation of human parsing within the framework of *Rational Analysis* (Anderson, 1991). Starting with the assumption that the human sentence processor is highly adapted to its task of analysing language quickly and correctly, they arrive at a probabilistic strategy that is mitigated by considerations of the cost of reanalysis. In order to parse correctly, the structural alterntive with the highest prior probability should be chosen, and in order to parse efficiently, the hypothesis that can be abandoned with least cost should be adopted.

The following is an overview over a range of mostly rule-based probabilistic models of human syntactic processing. The first probabilistic model of lexical access and syntactic disambiguation was proposed by Jurafsky (1996). The model uses a PCFG as a backbone. By a Bayesian approach, it evaluates the conditional probability $P(w|e)$ of the current word given the evidence already present in the system. Evidence can be top-down (probability information from some grammar rule) or bottom-up (information from a lexical entry). Lexical entries contain information about the preferred lexical category and subcategorisation frame of the item. Probabilities for these preferences as well as for the grammar rules are extracted from several corpora and norming studies. The top-down and bottom-up probabilities are directly combined by multiplication. This is not directly motivated by a well-studied formalism and has been a point for criticism.

The model works in parallel, constructing several possible alternative structures at a time. Structures that fall below a probability threshold are pruned from the beam of accessible interpretations. Effects of processing difficulty are modelled by showing that the correct analysis is not on the restricted beam of accessible interpretations at the moment of disambiguation, because it has been pruned out at some earlier stage.

[1] see Section 2.1.1: German PP attachment defaults to low attachment, but relative clause attachment defaults to high attachment.

The model relies on a context-free grammar backbone and therefore is hard to incrementalise, since a CFG only assigns probabilities to finished structures. Its performance is shown for hand-constructed examples only (no broad coverage). The model correctly accounts for the main clause/reduced relative ambiguity which can lead to serious parsing problems (Garden Paths) for sentences like *The horse raced past the barn fell*. The ambiguity arises because the verb *raced* is interpreted as a simple past, forming a main clause together with *the horse*. The following material makes this interpretation impossible and disambiguates towards a reduced relative clause modifying *the horse*, in which *raced* is a past participle. It also accounts for PP-Attachment through verb subcategorisation preferences.

Narayanan and Jurafsky (2001) carry on the main idea of Jurafsky's earlier work by constructing Bayesian belief nets to model human sentence processing. This is theoretically a much cleaner way of dealing with probabilistic evidence of different provenience and different nature than the approach in Jurafsky (1996). The model is again not a model of broad coverage, but is truly incremental because the probability for the alternative structures is re-computed with every new word from the input. The probabilities for sentence structure alternatives constructed by the belief net are able to account for human reading times in the case of the classic main clause/reduced relative clause ambiguity.

The following papers specifically address the problem of coverage by constructing models that are able to process normal corpus data with an acceptable accuracy as well as model phenomena of special interest.

Crocker and Brants (2000) propose an incremental system that consists of multiple layers of Hidden Markov Models. They model several well-studied phenomena in psycholinguistics (noun/verb lexical ambiguity, main clause/reduced relative ambiguity, NP/S-ambiguity). The Markov models are trained on corpus data to estimate the probability of chains of symbols given their input. Each of the layered models constructs all possible phrases over its input, given the phrase structure rules of a PCFG. The inital layer assigns word tags, later laters compute the most likely sequence of e.g. NPs, PPs and VPs over this input, and a final layer decides whether these can be combined into a sentence. Since each layer deals with chains of input symbols that are updated at every new input word, the cascaded Markov models are truly incremental. The probability information for this model is estimated automatically from one corpus, namely the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993), and the way probability estimates from lower levels are incorporated into higher levels is fully transparent and given by the mathematical theory of the model (as opposed to the treatment in Jurafsky (1996)). The model parses an unseen section of the Penn Treebank with an F-score that is lower than the current standard for parsing, but acceptable. It thereby demonstrates broad coverage.

Hale (2001) uses a fully parallel parser with a PCFG grammar. The parser is not a purpose built psycholinguistic model, but rather a standard application for probabilistic parsing. Although no coverage data is given, this sort of parser is explicitly constructed to parse large amounts of text, and the extensions used for modelling do not reduce that capability. Local processing difficulty in the face of ambiguity is modelled by comparing the probability of all possible syntactic structures that have been disconfirmed at the time of analysis (*word-by-word surprise*). A high value of surprise indicates that the current analysis is fairly unlikely and predicts processing difficulty. This is another theoretically clean way of achieving incrementality with a PCFG, since the sum over all possible outcomes starting from the current analysis is used. The model is evaluated on a single phenomenon, namely a Garden Path induced by the main clause/reduced relative clause ambiguity.

A problem for PCFGs in psycholinguistic modelling are phenomena that involve exactly the same grammar rules, only in a different order. In the final product of rule probabilities, the ordering does not show up, of course, so the structures are assigned the same probability although one may be preferred over the other by humans. One phenomenon that causes such difficulties in many PCFGs is the relative clause attachment ambiguity, as in *Who shot the servant of the actress who was on the balcony?* (Cuetos and Mitchell, 1988). The lack of global information on attachment preference in PCFGs that would allow the ambiguity to be resolved is an instance of the *Grain Size Problem* outlined in Mitchell et al. (1995), namely the difficulty of deciding on the correct level of syntactic analysis for the compilation of statistics.

Sturt et al. (2003) propose a hybrid model that overcomes the Grain Size Problem and that of a lack of training data in annotated corpora of limited size. The model uses a symbolic grammar formalism to encode sentence structure and a neural network to rank the resulting analyses. Since the network is free to pick its own evaluation criteria, this approach elegantly overcomes the Grain Size Problem. By generalising efficiently from seen to unseen cases, the neural network also tackles the problem of sparse training data.

The model relies on immediate integration of new material into existing structure at one of possibly several attachment points and is thereby purely incremental. To date, only syntactic factors are taken into account. The neural network is trained on corpus data from the Penn Treebank.

The model was tested on 500 sentences of unrestricted English text, also from the Penn Treebank. Its task was to predict the attachment of the next word given the correct syntactic analysis of the input so far. In 80% of cases, the model chose the correct attachment, and in 93% of cases, the correct attachment was within the top three choices. A wide range of phenomena from the psycholinguistic literature is modelled successfully, among them PP-Attachment by subcategorisation preference.

In a similar vein, McRae et al. (1998) use a competition-integration model (Spivey-Knowlton, 1997) to rate attachment decisions. Their model is not designed for broad coverage of language data. It also does not make claims about how the sentence structures to be rated are built.

The model is based on the idea of competing activations that stems from the neural network literature. Pre-defined constraints (information from the input) activate two nodes which correspond to the model's attachment decision. This is carried out in cycles until one of the nodes reaches a pre-defined threshold. If all incoming information favours one decision, the settling process is very short, but if there is conflicting information, it takes longer until one node reaches threshold. The number of cycles needed for each decision is assumed to be directly related to processing time in humans. McRae et al. (1998) successfully model the influence of thematic fit on the main clause/reduced relative ambiguity, as in *The cop arrested by the detective was guilty of taking bribes.* If the first NP is a good agent for the verb, readers prefer the main clause reading, showing difficulty at the disambiguation towards a reduced relative. If it is not, they take longer to process the initial part of the sentence until the reduced relative is disambiguated. Four constraints are used for this model. One is the thematic fit between the initial noun phrase and the agent and patient roles offered by the verb. The others regard the preference of the verb to appear as a simple past or past participle form, the bias introduced by *by*, and a general preference for the main clause over the reduced relative reading.

2.3 Disambiguation of PP-Attachments with Frequency Counts

The ambiguity of PP-Attachment is not only a potential problem for the human parser, but for parsing in computational linguistics as well. Because both verb subcategorisation and semantic influences determine its outcome, is difficult for a purely syntactical parser to find the correct attachment. The standard strategy to improve the correctness of PP-Attachment is to determine how often both attachment alternatives have been seen in corpus data and then to attach the PP according to the more frequent alternative.

The first such frequency-based approach of PP attachment disambiguation in parsing has been taken by Hindle and Rooth (1991), who used frequencies from a corpus to decide the attachment of phrase chunks. They counted the number of times the configurations of attachment site (verb or noun phrase) and preposition were seen in a corpus. These counts were combined into the *Lexical Association* ratio of the number of verb phrase attachments for the prepositon and attachment sites in question over the number of noun phrase attachments. This was also a way of approximating the attachment preferences for the verbal heads of possible attachment sites that were not accounted for in the grammar used for chunking. Deciding new attachment problems with the Lexical Association procedure resulted in 79.7% correct attachments. Ratnaparkhi (1998) has extended this approach to work in an unsupervised fashion by gaining attachment counts from unannotated corpora, while increasing the percentage of correct attachments to 81.9% for English (although on a different test corpus, so the figures are not directly comparable).

The approach of counting attachment configurations has also been fleshed out to take both possible attachment sites, the preposition and the head noun of the PP into account to decide the attachment. Different machine learning techniques have been used to train statistical models for this task (maximum entropy models, Ratnaparkhi and Roukos (1994) ; rule-based models, Brill and Resnik (1994)). All models suffer from the lack of training data, the *sparse data problem*, because the example attachments for the current ambiguity may not have been seen. There are different strategies to deal with this data sparseness, including the use of additional semantic information from ontologies.

The best-performing model for English that does not use semantic information is the one presented in Collins and Brooks (1995), reaching 84.5% correct attachments. This model makes use of configuration counts for all four words possibly involved in an instance of PP-Attachment (verb, object noun, preposition, noun in the PP). In case no counts have been seen for the exact configuration in question, the model *backs off* to using a combination of counts for just attachment triples, e.g. verb, preposition and the noun in the PP.

Attachment figures for the best model using semantic information are even higher at 88% (Stetina and Nagao, 1997). This model uses a semantic dictionary to define a measure of similarity between words so that similar words can be exchanged for one another in the quadruple sets to improve the number of counts of equivalent configurations. Ratnaparkhi and Roukos (1994) and Brill and Resnik (1994) use a similar strategy of bolstering sparse quadruple counts by generalising to semantic classes.

Another way of avoiding the problem of missing counts is to use a larger corpus. Work by Martin Volk on German (Volk, 2000, 2001) uses the World Wide Web to estimate co-occurrence counts for the configurations. He reaches correct attachment figures of 73% using a ratio of configuration trigram counts. To tackle remaining instances of missing counts, he also considers cases in which counts are available for one site only, as long as the counts are above some threshold. For the standard English data, this method only reaches 72% correct attachments (Lapata and Keller, 2003), so it is far worse than the current state-of-the-art model without semantic information.

Chapter 3

The Task

The models of human parsing presented in the previous chapter assume that human sentence processing can be modelled (at least to a large extent) by purely stochastic means. Their results seem to indicate that this assumption is indeed true. This thesis investigates the assumption for a new language (German) and a new syntactic phenomenon (PP-Attachment in verb final sentences).

3.1 The Data

The most reliable experimental results on PP-Attachment preferences in German come from Konieczny et al. (1997). Experiment 1 and 2 of this study are relevant here as they investigate PP-Attachment in verb second and verb final sentences. (Experiment 3 focuses on a different attachment phenomenon).

Experiments 1 and 2 use similar materials. Experiment 1 varies verb placement, subcategorisation preferences, and disambiguation of attachment through a semantic bias that renders one attachment more plausible than the other. Sentences 3.1 to 3.4 are example materials for each condition. Verb placement was varied between verb second and verb final position. Verb second sentences show an SVO word order just like English sentences (see sentence 3.1), while verb final sentences have an SOV order as in sentence 3.3. When the PP is encountered in SOV sentences, readers have to decide in the absence of the verb whether it is a modifier of the NP-object as in 3.3

or of the sentence predicate as in 3.4.

The verbs' subcategorisation preferences were for just an NP object (NP frame) or for both an NP and a PP object (NP-PP frame). The verb used in the examples is a verb with a preference for the NP-PP frame, so it would have a bias for attachment of the PP to the verb.

Semantic bias was varied by changing the noun of the PP to make it a plausible modifier of either the NP object or the verb. This bias decides the final outcome of the attachment. (See *with the skirt* in 3.1 and 3.3 *with the rock music* in 3.2 and 3.4)

(3.1) verb second, NP bias:

Iris	*störte*	*die Rentnerin mit dem Rock.*
Subject	Verb	Object
Iris	annoyed	the pensioner with the skirt.

Iris annoyed the pensioner with the skirt.

(3.2) verb second, Verb bias:

Iris	*störte*	*die*	*Rentnerin*	*mit*	*der*	*Rockmusik.*
Iris	annoyed	the	pensioner	with	the	rock music.

Iris annoyed the pensioner with the rock music.

(3.3) verb final, NP bias:

Neulich hörte ich,	*daß*	*Iris*	*die Rentnerin mit dem Rock*	*störte.*
		Subject	Object	Verb
Recently heard I,	that	Iris	the pensioner with the skirt	annoyed.

Recently I heard that Iris annoyed the pensioner with the skirt.

(3.4) verb final, Verb bias:

Neulich	*hörte*	*ich,*	*daß*	*Iris*	*die*	*Rentnerin*	*mit*	*der*	*Rockmusik*
Recently	heard	I,	that	Iris	the	pensioner	with	the	rock music

störte.
annoyed.

Recently I heard that Iris annoyed the pensioner with the rock music.

Experiment 2 varies only verb placement and semantic bias. Here, the PP is held constant while the verb and first NP are changed to vary semantic bias to exclude possible sources of noise from comparing reading times on two different PPs. Sentences 3.5 and 3.6 give an example of such an alternation.

(3.5) *Neulich hörte ich, daß Bruno den Jäger mit dem Gewehr*
Recently heard I, that Bruno the hunter with the rifle
fesselte.
bound up.
Recently I heard that Bruno bound up the hunter with the rifle.

(3.6) *Neulich hörte ich, daß Bruno den Hasen mit dem Gewehr erschoß.*
Recently heard I, that Bruno the hare with the rifle shot.
Recently I heard that Bruno shot the hare with the rifle.

The study was conducted by tracking the eye movements of subjects while they were reading both SVO sentences and SOV sentences. Significant effects were found on the noun of the PP. They show up most reliably for Regression Path Durations. This measure accounts for the amount of time spent on a region and preceding regions until the first forward eye movement. To exclude contamination by unrelated sentence-final processing and eye movement effects, only the first re-readings of prior regions were considered whenever the PP was the last region of the sentence. Regression Path Duration is a *late* measure, because it considers more than just the time spent on initial reading of a region. This means that it can detect re-readings induced by processing problems, too.

The results from Experiment 1 are that reading times at the PP for verb-final sentences with a verb attachment bias is longer than for verb-final sentences with NP-Attachment bias. Also, by subjects, there is an effect for verb-second sentences such that for NP frame verbs, reading times are increased when the object is biased towards verb attachment, while the opposite holds of verbs with an NP-PP frame. Experiment 2, which was run with NP-PP verbs only, replicates this result for verb-second sentences both by items and by subjects. Also, there was an effect by subjects on the verb of verb-final sentences, namely an increase in reading times if the PP was biased towards NP-Attachment.

Taken together, these results provide evidence that, for verb-final sentences, there is a preference to initially attach the PP to the NP (effect 1 from Experiment 1 and by subjects effect from Experiment 2). In verb-second sentence, where both attachment sites and the verb's attachment preferences are known, attachment is preferentially with the verb's preference (by subjects result from Experiment 1 and effect 1 from Experiment 2).

3.2 The Task

This Thesis further investigates the hypothesis that human sentence processing can be accurately modelled using frequency information from large amounts of text. The task is to build a stochastic model of sentence processing that accounts for the robustness of the human language processor as well as for the attachment preferences found in the Konieczny et al. (1997) study. The effects of verb subcategorisation preference for verb second sentences have already been modelled for English (Jurafsky (1996); Crocker and Brants (2000); Sturt et al. (2003)). The condition of most interest are verb-final sentences, as there are to date no statistical models for the incremental processing of these sentences.

If a statistical model can be shown to account for the effects at the PP in both the verb second and verb final conditions, the position of frequency based models of human parsing would be strengthened further. If one of the effects is shown not to be covered by the model, this would indicate a limitation of the widely used statistical accounts.

3.3 The Architecture

The task outlined above will be tackled by a stochastic parser. Its broad coverage of corpus data accounts for the broad variety of language phenomena that humans process without apparent difficulty. The simplest possible stochastic parsing model (see Section 4.4) only covers the verb placement variable adequately. Verb subcategorisation information has to be explicitly incorporated into the grammar. The parser also has to report the intermediary state of its analyses at the noun of the PP to allow an analysis of its incremental parsing decisions.

On top of this parser, the influence of semantics can be stipulated because for each condition, it is known which attachment preference prevails. A conflict between the parser's decision and the semantic bias of the condition can then be interpreted as predicting longer reading times because reanalysis or a re-ranking of structure has to take place. We also attempt to approximate semantic decisions by a shallow semantic module. This makes the final attachment decision on the basis of co-occurrence counts of the noun in the PP and the two attachment sites. Again, a conflict between this decision and the parser's indicates processing difficulty.

Chapter 4 describes in detail the development of the parser-based model, while Chapter 5 goes into the add-on semantic module. Chapter 6 presents the results on a previously unseen test set for each module separately and for the model as a whole.

Chapter 4

Module I: Syntactic Information

4.1 Overview

The backbone of the syntactic model is the stochastic left-corner parser LoPar (Schmid, 2000), which is described in Section 4.2. As outlined in the last chapter, this parser is extended with information about the subcategorisation preferences of verbs and then modified to output the state of its analysis at the PP during the processing of a sentence.

The following sections describe the different elements of the syntactic module. Section 4.3 describes the training and testing data and Section 4.3.1 investigates general properties of that data, such as attachment preferences and subcategorisation preferences evident in it. Section 4.4 gives an overview over the performance of the Baseline model, which consists of the parser and the simplest possible grammar with no added information. Section 4.5 describes the incorporation of subcategorisation information, Section 4.6 goes into the treatment of several sparse data phenomena and Section 4.7 treats the adaption of the parser to output preliminary results after processing the PP.

4.2 The Parser

LoPar is an integrated tagger and parser. It uses raw text as input, determines the optimal sequence of part of speech tags and builds all structural analyses of the input sentences that are licensed by the grammar. This is done in a fully parallel way, so all possible analyses are generated and stored concisely in a chart.

LoPar is a stochastic parser which assigns structures to input sentences on the basis of a probabilistic CFG. It implements the *Left-Corner* parsing strategy. This parsing strategy circumvenes the problems of both pure top-down and pure bottom-up parsing. Pure top-down parsing does not take the input string into account at all while phrases are predicted from the desired goal phrase (e.g. *S*) downwards. The input string is only considered at the very last step when the predicted word categories have to be linked to input words. Wrong predictions can therefore only be corrected at the very last step of parsing, which means that unnecessary work is done. Pure bottom-up parsing starts from the input, combining the known word tags into phrases and those into higher-level phrases until the goal phrase has been constructed. This aimlessly generates all possible combinations of the input word categories, which is again be labour-intensive for a large grammar.

The left-corner strategy combines top-down information from the phrase structure rules and bottom-up information from the input through lexical rules. This ensures that there is a top-down constraint on which phrases should be built next to complete the structure in question while the bottom-up information from the input string is used right away. This ensures that predictions which cannot be borne out by the input are abandoned quickly. A left-corner parser will alternate bottom-up and top-down steps and try to match top-down predicted phrases to bottom-up constructed ones. This linking is done by the left corner of the grammar rules, that is the leftmost of the right-hand side non-terminals. As soon as a bottom-up step completes a non-terminal node (a word tag or a phrase), this is matched to the left corners of all known grammar rules. If the left-hand side of any of those rules matches a top-down prediction, the current phrase is connected to the predicted phrase as a daughter and the link is made. Even if there is no link, the rules which match the left corner can be used to predict new local goal phrases to complete the phrase licensed by the rule. Through this interaction of prediction and linking up from the left corner of a rule, the Left-Corner strategy produces a full syntactic tree with root node and words at the leaves from the first word of the input, unlike the Top-Down strategy, that includes the actual words last, and the Bottom-Up strategy that generates the tree root last. The Left-Corner tree is then extended with every word in a quasi-incremental way.

4.3 Materials

Two sets of sentence materials and a lexicon of subcategorisation information are used to build the model. To train the syntactic module and evaluate its coverage, the NEGRA corpus (Skut et al., 1997) is used. This is a 20,600 sentence corpus of German newspaper text (355,000 tokens). The sentences in the corpus are annotated with constituent structure and some grammatical functions (head, subject, object, modifier). The annotation scheme assumes flat structures in order to be able to cope with freedom of word order effects 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. This means that scrambling phenomena just alter the sequence of sisters in the tree, while otherwise they would require complex usage of traces. The annotation scheme also allows crossing branches in syntax trees to allow discontinuous constituents. In the treebank format used here, these crossing branches are replaced by trace and filler markers.

The first 18,600 sentences of the corpus are used as training data. The next 1000 sentences make up the development set and the last 1000 sentences form the test set. All traces were removed from the corpus, because PCFGs assume independence between rule applications and cannot deal with the relationship between filler and trace in a meaningful way. Grammatical function labels were also removed. Finally, to improve parsing efficiency, sentences with more than 40 words were removed from all three sets. This reduced the development set to 975 and the test set to 968 sentences. The training set was reduced to 18,000 sentences.

The subcategorisation lexicon for German words compiled by Schulte im Walde (2002) is used to bolster sparse counts for verb subcategorisation frames. She extracted frame counts for c. 17,000 different verbs from a large newspaper corpus. Evaluation against a hand-written standard dictionary of verb usage established its data as fairly reliable.

The second set of sentence materials consists of the 160 experimental items from Experiments 1 and 2 of the Konieczny et al. (1997) paper. These were split into a test set and a development set, which is used to estimate the performance of the syntactic and semantic modules on sentences of the same type as the test sentences. This is necessary because the experimental items differ noticeably from the NEGRA data in style, so good performance on the NEGRA test set does not automatically imply good performance on the experimental items and vice versa.

A development and a test set were compiled from the full set of experimental items. First, we deleted four sentences which contain the one verb which is not accounted for in the either the NEGRA corpus or the subcategorisation lexicon (*zerknicken (crumple, bend)*). From the remaining 156 sentences, five sentences in each condition of Experiment 1 and seven of each condition of Experiment 2 were chosen randomly, to form the development set of 68 sentences. The remaining 88 sentences form the test set. The sentences in the development and test sets can be found in the Appendix.

4.3.1 Pretests

Since the performance and predictions of any stochastic model depend crucially on the nature of the training data, we report two analyses of important aspects of the data. One concerns the overall PP-Attachment preferences in the NEGRA corpus, and the other the verb subcategorisation preferences evident in the NEGRA corpus and the subcategorisation lexicon for German which was used to bolster sparse NEGRA counts (see Section 4.6.1).

A test of PP-Attachment preferences in the NEGRA corpus shows that of the 6261 instances of PP-Attachment where both an NP and a verb are present, 38.2% of attachments are to the noun and 61.8% to the verb. The NEGRA corpus therefore reflects a preference for PP-Attachment to the verb rather than the noun phrase. This is possibly not generally true for German, since Volk (2001) reports 63% noun attachment in his ComputerZeitung test corpus (out of 4383 PP-Attachment constructions).

As an additional source of information about verb subcategorisation preferences, the subcategorisation lexicon for German compiled by Sabine Schulte im Walde is used (Schulte im Walde, 2002). An analysis of verb preferences in that lexicon and in

NEGRA showed that the subcategorisation preferences for the verbs differ noticeably from those established in completion studies for the Konieczny et al. (1997) materials. We used the measure from Garnsey et al. (1997) to determine verb bias towards one subcategorisation frame or the other: A word is classified as being biased for an NP and a PP object rather than just a single NP object if it appears more than twice as often with both a PP and an NP object than with just an NP object. For the materials from Experiment 1, five verbs out of the 12 items in the PP-subcategorisation condition were attested in NEGRA. Out of those, four were biased towards taking just an NP object, and the same was true for nine out of the 12 verbs in the subcategorisation lexicon. In the NP-subcategorisation condition, which also has 12 items, five out of eight attested verbs in NEGRA were biased towards taking an additional PP object, while four out of the 11 verbs attested in the lexicon showed this bias and the remaining seven showed no marked preference for either frame. For the 30 items from the PP-subcategorising materials in Experiment 2, again the majority of verbs in the NEGRA corpus and in the lexicon were biased towards just an NP object (NEGRA: 10 out of 16, 5 PP-biased; lexicon: 26 out of 30, 1 PP-biased).

However, these preferences are not necessarily directly mirrored in the model's performance. The rules of a PCFG work top-down, that is in the $POS \rightarrow word$ direction, so their probabilities are determined by dividing the frequency with which *word* occurs as *POS* by the total frequency of *POS*. Since the NP frame is more frequent in general than the NP-PP frame, a larger raw count of the NP frame does not necessarily lead to a substantially larger probability for this frame than for the NP-PP frame. In order to ascertain the biases as they appear in the model, Garnsey et al.'s test was again applied to the lexical rule probabilities in the model once the frequencies from NEGRA and the subcategorisation lexicon had been combined. The comparison was made between the probabilities for the rules $NP\text{-}frame \rightarrow word$ and $NP\text{-}PP\text{-}frame \rightarrow word$. There are indeed more equibiased verbs when rule probabilities in the model are considered, but in each condition, more than half of the verbs still show a reversed subcategorisation bias. Out of the 12 supposedly NP-PP-subcategorising verbs in Experiment 1, seven are biased towards NP-subcategorisation and five show no bias. In the NP-subcategorisation condition, six out of eleven attested verbs are biased towards NP-PP-subcategorisation, four show no bias, and only one prefers a single NP object. In Experiment 2, 16 out of 30 verbs still show an NP-subcategorisation bias, with 11 unbiased verbs and only three showing a clear bias towards taking a PP object. In sum, it has to be expected that the model's predictions for the subcategorisation preferences will be the opposite of the preferences in the Konieczny et al. data in most cases. At every step of development of the syntactic module, we will analyse whether this prediction is really borne out in the module's performance.

The equivalence of attachment preferences from corpora and production studies has been a matter of discussion in the field. Merlo (1994) and Gibson et al. (1996) provide evidence against a positive correlation between the data sources. Roland and Jurafsky (2002) argue that verb subcategorisation frequencies differ between corpora of psycholinguistic sentence production data, written discourse and conversation data due to the influence of discourse type and verb sense that are specific to corpora. For the British National Corpus (BNC), which is a balanced corpus made up from spoken and written language, Lapata et al. (2001) have indeed shown a reliable correlation of corpus and completion data. Our results again yield no evidence for a reliable positive correlation of corpus and completion data, but the corpora used here both contain newspaper text only, so they possibly do not reflect general language usage well.

4.4 Baselines

As a first step towards building and evaluating the syntactic module, a Baseline is provided as a lower bound on performance which serves as a comparison for further development steps. To provide a Baseline for the tasks at hand, the parser was equipped with an unlexicalised grammar induced from the training section of the NEGRA corpus. The Baseline grammar contains all the rules that can be read off the tree structures of the training section of the corpus and their frequencies. A lexicon of word forms and their frequency was also induced from the corpus.

LoPar is an integrated parser and tagger, which means the algorithm determines the optimal sequence of tags for the input and the optimal parse at the same time. Of

Condition	Precision	Recall	F-Score	Tagging Accuracy	Coverage
Baseline	71.30	72.45	71.87	95.27	99.2
Baseline, PT	73.51	75.55	74.51	-	96.9

Table 4.1: Baseline results on the NEGRA development set

course, inaccuracies in the predicted tag sequence, e.g. when words are not in the lexicon and cannot be tagged, can diminish the performance of the parser proper. In order to provide an upper bound of the performance of the parser proper, the correct tags for the words in the development set were provided in a second run. In this case, LoPar accepts the tag sequence as given and just constructs the optimal syntactic structure over it.

Table 4.1 shows the performance of the Baseline model on the NEGRA development set. We report Labelled Precision, Labelled Recall and F-Score ($\frac{2 \times Precision \times Recall}{Precision + Recall}$). Labelled Precision is the number of correctly labelled phrases with correct span that the parser found by the number of all phrases it assigned. It measures how reliable the parser's phrase assignments are. Labelled Recall is the number of correctly labelled phrases assigned by the parser by the number of phrases in the hand-annotated version of the input corpus. This measure indicates how well the parser is doing in comparison to the ideal structures. In addition to these measures, we give the accuracy of the parser's tagging and its coverage, that is the percentage of sentences that it assigned structure to.

The Baseline model is able to assign structure to 99.2% of the 975 sentences in the development set. The F-Score for these structures is 71.87, and 95.27% of all tags are correct. When Perfect Tags (PT in tables) are provided, the F-Score rises to 74.51, which is an upper bound on the performance of the parser proper with the Baseline grammar. Coverage drops to 96.9% when Perfect Tags are used, because some of the ideal tag sequences have not been seen in the training corpus and there is no rule in the grammar that would license combining them into a sentence structure. When the tags are not fixed, the input can still be processed at a cost to F-score and tagging accuracy by assigning some other, less probable tag to some of the input words.

Condition	Long Items	Shortened Items	out of
NP-PP frame, verb final	6, NP	5, NP	10
NP-PP frame, verb second	10, NP	10, NP	10
NP frame, verb final	4, NP	5, NP	10
NP frame, verb second	9, NP	10, NP	10
NP-PP frame, verb final	11, NP (2 V)	12, NP	14
NP-PP frame, verb second	12, NP	14, NP	14
Accuracy	79.4%	82.4%	
Coverage	100%	100%	
Correct Attachments	27.8 (72.2) %	26.8 (73.2)%	

Table 4.2: Baseline results on long and shortened experimental items: Attachment predictions per condition, overall Accuracy, Coverage and the percentage of correct attachment predictions assuming Konieczny et al.'s preferences (or the ones evident in our data)

The experimental items prove to be rather difficult to parse with the Baseline model because they contain additional material after the PP to facilitate the detection of effects in the eyetracker experimental setting (the so-called *spillover region*). This makes the sentences rather unlike the NEGRA data and therefore results in bad performance of the Baseline model, as shown in Table 4.2. In this table and all subsequent ones, the first four conditions of experimental items stem from Experiment 1, the last two from Experiment 2. Since there is no hand-annotated version of the experimental items for comparison, we report Coverage (the percentage of sentences assigned some structure) and Accuracy (the percentage of correctly parsed sentences out of all input sentences). Sentences count as being correctly parsed if the phrase structure is correct in the clause in question (errors in enclosing clauses are ignored, as are failures to close the clause off properly as long as the attachment is clear). Also, the phrase tags have to be correct for the verb, NP and PP (for verbs, any verb tag is admissible, e.g. auxiliary instead of full verb).

All sentences are assigned some structure, but only about three quarters of the assigned structures are the intended ones, so accuracy is only 79.4%. Since the additional material is in itself of no interest to the experiment or the modelling of the effects reported, it was removed to leave much shorter, more standard sentences that are easier to parse. Adjectives in the PPs were also removed for the same reason. The Baseline results on these shorter items are also presented in Table 4.2. For these sentences, accuracy reaches 82.4%, while coverage remains at 100%.

From a modelling perspective, not only the number of correctly parsed sentences is interesting, but also the percentage of correctly parsed sentences that show the correct attachment decision. For the syntactic module, attachments count as correct if they mirror the syntactic subcategorisation preference of the condition. In the absence of a semantic module, the parser cannot be expected to account for the semantic disambiguation as well. As can be seen in Table 4.2, the Baseline parser almost always chooses NP-attachment on the original items. NP-Attachment is indeed the psycholinguistically correct default for German (Konieczny and Hemforth, 2000). The Baseline is therefore almost identical to the theoretical Baseline of always choosing the default attachment. The percentage of correct attachments over all correctly parsed items is 27.8% for the subcategorisation preferences assumed in Konieczny et al. (1997). The theoretical Baseline assigns the correct attachment to 29.4% of all sentences, namely the 20 sentences in the NP-frame conditions. Since the verb subcategorisation preferences of the data used here are probably the exact opposite of those preferences, it has to be assumed that three quarters of the experimental items preferentially subcategorise for just an NP in our model, so the Baseline's predictions are correct in 72.2% of all correctly parsed cases.

On the short items, the Baseline predicts only NP-Attachment. The predictions remain essentially the same, however (in absolute numbers, there is a difference of one attachment prediction between the runs on long and short items).

The performance of this Baseline serves as a standard of comparison for the performance of more elaborate versions of the model.

4.5 Subcategorisation Information

The first improvement over the Baseline model is the addition of verb subcategorisation preferences to the grammar. This allows the parser to take these preferences into account when making attachment decisions. The preferences are added to the grammar by extending the set of verb tags. The new tags distinguish between verbs with different subcategorisation preferences, e.g. NP and NP-PP. To arrive at such differentiated verb tags, every instance of the original STTS (Schiller et al., 1995) verb tags in NEGRA is annotated in the training corpus with the argument structure the verb appears with in that sentence. The tags themselves already distinguish between auxiliary, modal and full verb and verb mode (that is, the infinitive, participle, imperative and finite form).

The argument structures each verb appears with are determined heuristically by counting the sisters of the verb as complements. Word order was not taken into account because of the relative freedom of German word order as evident in verb final sentences and scrambling phenomena. Traces are ignored as complements, because they cannot be handled by the context-free grammar LoPar uses and therefore cannot be included in the input to the parser. Phrases that are marked as displaced from another head and phrases that are marked as adjuncts were excluded from consideration. Unfortunately, only noun and verb phrases are explicitly marked as complements or adjuncts in NEGRA. All prepositional phrases therefore have to be considered as potential complements. However, they are annotated as complements in a conservative way only if the verb appeared with a PP and a single NP (NP-PP frame) or no complements other than a PP (PP-frame). The numbers for the NP-PP frame are probably still slightly overestimated. The same is true for the the null and NP frames because traces were not considered.

The verb frames chosen for annotation are a conflation of the ones identified by Sabine Schulte im Walde's subcategorisation lexicon for German (Schulte im Walde, 2002). As an example of collapsed distinctions, the lexicon treats reflexives separately from normal noun phrases, while here, they are counted as noun phrases. This is possible because many reflexive pronouns just fill an existing NP argument slot of a verb, as in German *sich waschen* (*wash*). Also, cases where the reflexive is lexical and does not fill a semantic slot as in *sich fürchten* (*be afraid*) cannot be discerned from cases of non-lexical reflexives in the NEGRA annotation. Also, case distinctions between objects made by the lexicon are collapsed and distinctions between expletive and full subjects are ignored here.

Apart from the frames already described in (Schulte im Walde, 2002), there were no additional frames worth of consideration in the NEGRA corpus. The final frames used for annotation describe the number of arguments other than the subject, which is marked up in NEGRA and can therefore be easily identified in the annotation process. The frames are *null, n, p, i, s, nn, np, ni* and *ns*. The *null* frame describes an intransitive verb, the *n, p, i* and *s* frames describe a transitive verb with a NP, PP, infinitive and sentence object, respectively. The ditransitive tags similarly stand for a combination of a noun phrase object with other NPs, PPs and infinitive or sentence objects.

Once the existing verb frames have been annotated with the frame tags, a second grammar can be read off the annotated corpus. This grammar incorporates subcategorisation information by allowing every verb to show its preferences for different frame sets through the verb tags it frequently appears with in the training corpus. These frequencies are used to calculate the probabilities of the lexical rules of the form $POS \rightarrow word$, which are used to find the most probable tag for a word in the input. A verb's tag in turn causes the choice of a sentence rule with the correct complement configuration. This allows the verb preferences to influence the attachment of the PP, which was impossible in the Baseline condition.

Condition	Precision	Recall	F-Score	Tagging Accuracy	Coverage
Baseline	71.30	72.45	71.87	95.27	99.2
SC	68.54	70.54	69.52	91.94	96.4
SC, PT	75.28	78.31	76.76	-	93.5
Baseline, TnT tags	70.13	73.19	71.62	96.76	96.0
SC, TnT tags	61.31	67.02	64.03	93.55	84.5

Table 4.3: Results for the grammar with subcategorisation information (SC) on the NEGRA development set with different input tags

When the grammar with subcategorisation information is used on the experimental items development set, both the F-Score and the tagging accuracy drop from the Baseline condition (see Table 4.3). The F-Score decreases to 69.52, and tagging accuracy goes down from 95.27% to 91.94%. Coverage also suffers. The Perfect Tag upper bound shows a potential rise in performance over the Baseline of similar strength as the drop in actual performance (F = 76.76). The reason for both this increase and the actual drop is that the number of possible tags has doubled in size due to the proliferation of verb tags. The original twelve verb tags have been expanded to a potential 108, out of which 81 are attested in the corpus. This leads to data sparseness and ambiguity in tagging.

Experimentations with an external tagger (TnT, Brants (2000)) as a pre-processing step to provide correct tags also fail to improve the results for this reason. The F-score for the Baseline model is not affected in either way by the use of TnT assigned tags, but the performance of the version with differentiated verb tags decreases steeply, along with coverage, while tagging accuracy increases from 91.94 to 93.55. The tagger appears to assign more tags correctly than the parser does, but the incorrect ones are evidently more damaging to the correctness of the phrases built on them. The new verb tags are the likely source of the error, because no other changes have been made to the tag set that the Baseline grammar uses. At the heart of the decline in performance is the sparse data problem: Not enough instances of the 81 verb tags have been seen to guarantee correct behaviour of the TnT tagger. Incorrect verb tags then cause incorrect sentence rules to be chosen, which leads to the decline in general performance. Also, coverage declines steeply as for many tag sequences, no sentence structures are licensed by the grammar rules.

On the experimental items development set, the percentage of correct attachments, coverage and accuracy also drop slightly. The results are summarised in Table 4.4. Inspection of the parser's output lets us identify three main causes of error: Performance drops partly because not every verb in the development corpus is attested with an appropriate frame in the lexicon. To alleviate this problem, additional data from the subcategorisation lexicon will be used (see Section 4.6.1). Also, many nouns in the PPs are rare compound nouns which are not listed in the parser's lexicon and which are therefore mistagged. This means that no correct PPs can be formed and consequently, accuracy suffers. The treatment of complex nouns is described in Section 4.6.2. Lastly, some of the grammar rules that would be necessary to build acceptable structure for the verb final sentences are missing. This problem and its solution is discussed in Section 4.6.3 below.

Condition	Baseline	Subcategorisation	out of
NP-PP frame, verb final	5, NP	4, NP	10
NP-PP frame, verb second	10, NP	9, NP	10
NP frame, verb final	5, NP	4, NP	10
NP frame, verb second	10, NP	7, NP (2, V)	10
NP-PP frame, verb final	12, NP	9, NP	14
NP-PP frame, verb second	14, NP	10, NP	14
Accuracy	76.5	66.2	
Coverage	100	97.1	
Correct Attachments	26.8 (73.2)%	24.4 (75.6)%	

Table 4.4: Results for the grammar with subcategorisation information on the experimental items development set

4.5.1 Lexicalisation of the Annotated Grammar

By extending the verb tags, subcategorisation preferences are recorded per verb class. A way of giving the parser more fine-grained information about subcategorisation preferences and even semantic attachment preferences is to lexicalise the grammar. A lexicalised grammar annotates every phrase with its lexical head. This way, verb specific subcategorisation preferences are recorded. Even selectional restrictions on the arguments of a verb can be modelled in a very preliminary way because the lexical arguments that a verb usually takes are also recorded. Lexicalised grammars are very big, however, because every new configuration of lexical heads generates a new rule, and they do not generalise well to unseen configurations. This is why LoPar, like most parsers, backs off to using unlexicalised grammar rules if no lexicalised rule is applicable. For English, lexicalisation usually leads to a 10% increase in performance, while for German, it has been shown not to be of much help for the same training data and parser configuration that was used here (Dubey and Keller, 2003). This is why lexicalisation of the grammar without subcategorisation information was not attempted. Lexicalising the grammar now might yield an even more fine-grained picture of subcategorisation preferences per word. This would improve performance on the experimental items. However, the introduction of a great number of possible verb tags already lead to sparse data problems for the unlexicalised grammar, so it is to be expected that lexicalisation of a grammar which contains all these tags will encounter an even greater sparse data problem. We did, however, produce a lexicalised version of the grammar with subcategorisation information.

For lexicalisation, the lexical head of every grammar rule needs to be known. In NEGRA, heads are annotated for the categories S, VP, AP (adjective phrase) and AVP (adverbial phrase). For the other phrases, heads were heuristically determined as in Dubey and Keller (2003). This is standard practice on the Penn Treebank, a widely used corpus of English with syntactic annotations (Marcus et al., 1993).

Performance on NEGRA indeed diminishes noticeably for the lexicalised version of the grammar with subcategorisation information. This is true both with and without Perfect Tags (see Table 4.5). The F-score without Perfect Tags is 59.42, which becomes 71.96 with Perfect Tags. The upper bound for parser performance with a lexicalised grammar barely rises above the unlexicalised Baseline of F = 71.87 and stays well below the upper bound performance for the Baseline of F = 74.51. Coverage stays at a level with the Baseline, while it had also dropped for the unlexicalised grammar with subcategorisation information.

Condition	Precision	Recall	F-Score	Tagging Accuracy	Coverage
Baseline	71.30	72.45	71.87	95.27	99.2
SC, lexicalised	56.21	63.02	59.42	85.64	99.3
SC, lexicalised, PT	68.29	76.05	71.96	-	92.6

Table 4.5: Results on the NEGRA development set for the lexicalised grammar with subcategorisation information

Condition	Subcategorisation	Lexicalised	out of
NP-PP frame, verb final	4, NP	1, NP	10
NP-PP frame, verb second	9, NP	8, V (1 NP)	10
NP frame, verb final	4, NP	0	10
NP frame, verb second	7, NP (2, V)	4, V (2 NP)	10
NP-PP frame, verb final	9, NP	2, V (1 NP)	14
NP-PP frame, verb second	10, NP	5, V (4 NP)	14
Accuracy	66.2	41.2	
Coverage	95.6	97.1	
Correct Attachments	24.4 (75.6)%	67.9 (32.1)%	

Table 4.6: Results on the experimental items for a lexicalised version of the grammar with subcategorisation information

On the experimental items, accuracy also drops dramatically to only 41.2%. Table 4.6 also gives the attachment predictions per condition and the coverage figures. Coverage actually increases slightly, mirroring the good coverage the lexicalised grammar achieves on the NEGRA development set. The percentage of correct attachments rises steeply if Konieczny et al.'s subcategorisation preferences are assumed and drops if the reversed preferences are assumed because there is a preference for verb attachment in all conditions. Since the lexicalised grammar shows so little accuracy even on the experimental items, it is not considered for use as the final model.

4.6 Sparse Data Handling

There are three evident problems for the module that have been caused by a lack of training data. One is the lack of reliable subcategorisation information for the verbs in the experimental data training set. Another is the difficulty the module has in dealing with the many rare nouns in that data set that are not listed with a part of speech tag in the parser's lexicon. And finally, some of the rules necessary to correctly parse the data do not exist in the grammar because a comparable sentence structure has not occurred in the training corpus. The solutions to these problems are now described in turn.

4.6.1 Sparse Subcategorisation Information

46% of the verbs in the development set are unseen in the NEGRA corpus, and those that are accounted for in the parser's lexicon often have not been seen with the two frames that are of interest here, namely the NP-PP and the NP frame. Therefore, additional verb subcategorisation information from Schulte im Walde's subcategorisation lexicon is used.

Condition	Precision	Recall	F-Score	Tagging Accuracy	Coverage
Baseline	71.30	72.45	71.87	95.27	99.2
SC, words	68.54	70.54	69.52	91.94	96.4
SC, words, PT	75.28	78.31	76.76	-	93.5
SC, lemmas	63.61	64.69	64.15	85.83	99.4
SC, lemmas, PT	75.31	78.31	76.78	-	92.6

Table 4.7: Results on the NEGRA development set for lemmatised input as opposed to words

Lemmatisation The subcategorisation information in the lexicon is available per lemma, but not per word form, so the words in the input data and the lexicon have to be lemmatised. This is done with the DMM lemmatiser (Lorenz, 1996). Lemmatisation by itself should improve results, because the possibly sparse counts for several morphological forms are conflated into a more reliable count for the lemma. However, as the results on the NEGRA corpus in Table 4.7 show, performance drops to F = 64.15 from 69.52. This is probably caused by the bad tagging accuracy which goes down steeply to 85.83% from 91.94% when the input consists of lemmas. A run with Perfect Tags confirms that the decline in performance is caused by the additional uncertainty about the correct tag that appears when counts for forms of a word, e.g. of a verb, are conflated to give a choice between finite, infinitive, imperative and participle form. This problem becomes more serious because there is also a choice of different subcategorisation frames. When Perfect Tags are provided, the word and the lemma condition show identical performance.

On the experimental items, the version using lemmas achieves less accuracy than the one using words, but its predictions are more correct assuming that the subcategorisation preferences per condition are reversed in comparison to those in Konieczny et al. (1997), which is in keeping with the preferences established for our data (see Table 4.8). There are first indications that this assumption is indeed true: In the NP-PP subcategorising conditions, more attachments are made to the NP than to the verb, and in the NP-subcategorising verb second condition, the attachment bias is clearly to the verb. Coverage also increases slightly, which is in keeping with the results on the NEGRA development set.

Condition	Words	Lemmas	out of
NP-PP frame, verb final	4, NP	4, NP	10
NP-PP frame, verb second	9, NP	8, NP (2, V)	10
NP frame, verb final	4, NP	2, NP (1 V)	10
NP frame, verb second	7, NP (2, V)	8, V	10
NP-PP frame, verb final	9, NP	5, NP	14
NP-PP frame, verb second	10, NP	8, NP (5, V)	14
Accuracy	66.2	63.2	
Coverage	95.7	97.05	
Correct Attachments	24.4 (75.6)%	20.9 (79.1)%	

Table 4.8: Results on the experimental items development set for words and lemmas as input

Combination of Frequency Counts To realise the addition of subcategorisation information to the parser's lexicon, the subcategorisation data from Schulte im Walde (2002) are added to the tag counts for each word from the NEGRA corpus. This increases the number of frames attested per verb and also the number of times each verb is attested with each frame, which makes the probabilities attached to the lexical rules more reliable. To achieve this, the data from the subcategorisation lexicon are converted to correspond to the conflated frame definitions used here and then combined with the existing frequencies from the NEGRA corpus. The frequencies from both resources cannot just be added up, however. That would distort the overall frequency distribution of tags because the frequencies from the subcategorisation lexicon come from a much larger corpus and are therefore larger than the NEGRA frequencies. Instead, the frequencies are combined by converting the raw co-occurrence frequencies of verbs and tags to percentages of the total frequency of that tag for each corpus separately. These percentages correspond to lexical rule probabilities as predicted by each corpus. These are then combined using a weighted average. A weighting factor of 0.5 amounts to averaging over both sources of information, while other factors give more weight to one of the sources rather than the other.[1]

Below, $P(POS \rightarrow word)$ is the rule probability of the rule that rewrites POS as $word$. f_o and f_s are the original and the supplementary rule frequencies, and w is the weighting factor.

$$P(POS \rightarrow word) = \frac{w * f_o(POS \rightarrow word)}{\sum_{word} f_o(POS \rightarrow word)} + \frac{(1-w) * f_s(POS \rightarrow word)}{\sum_{word} f_s(POS \rightarrow word)}$$

The above computation results in probabilities for the $POS \rightarrow word$ rules, while the parsers expects tag frequencies in the lexicon and computes the probabilities itself. Therefore, the right hand side of the equation is brought to the common denominator, which can then be dropped to yield frequencies. The final formula thus becomes

$$\begin{aligned} f(POS \rightarrow word) \ = \ & w * f_o(POS \rightarrow word) * \sum_{word} f_s(POS \rightarrow word) + \\ & (1-w) * f_s(POS \rightarrow word) * \sum_{word} f_o(POS \rightarrow word) \end{aligned}$$

Results The use of the additional subcategorisation information increases the F-Score on the NEGRA development set substantially over the initial subcategorised model using lemmas (F = 67.68), but the new model does not achieve the level of

[1] The weighting factor w was varied from 0.3 to 0.9, but this had only negligible effects on both the performance on NEGRA and on the experimental items.

Condition	Precision	Recall	F-Score	Tagging Accuracy	Coverage
SC, words	68.54	70.54	69.52	91.94	96.4
SC, words, PT	75.28	78.31	76.76	-	93.5
SC, lemmas	63.61	64.69	64.15	85.83	99.4
SC, lemmas, PT	75.31	78.31	76.78	-	92.6
SC, add. counts	67.11	68.28	67.68	89.08	99.5
SC, add. counts, PT	74.39	77.26	75.79	-	92.8

Table 4.9: Results on the NEGRA development set for the grammar with additional subcategorisation information

performance of the initial subcategorised model with full word forms (see Table 4.9). The Perfect Tag upper bound also demonstrates a small decrease in potential performance over the initial model using just words. Coverage rises to 99.5%, which slightly surpasses baseline coverage.

Table 4.10 shows that on the experimental items, accuracy reaches the same level again as for full word forms. The percentage of correct attachments even rises above the full word form condition, although not reaching the performance of lemmas only. Coverage is perfect again for the first time since amendments were made to the Baseline model. The increase of performance on the experimental items is more important for the model than the slight drop in performance on the NEGRA data.

4.6.2 Rare Compound Nouns

Another notorious cause of errors on the experimental items is that many of the composite nouns in the experimental items are infrequent and therefore mistagged. This upsets the parsing process. For example, *Füllfederhalter* (*pen*) or *Brandzeichen* (*brand*) are tagged as adjectives because the parser assigns the most frequent tags to unknown words. Consequently, the PPs cannot be formed correctly and the whole sentence is misparsed.

Condition	Words	Lemmas	Additional Counts	out of
NP-PP frame, verb final	4, NP	4, NP	3, NP	10
NP-PP frame, verb second	9, NP	8, NP (2, V)	8, NP (2, V)	10
NP frame, verb final	4, NP	2, NP (1 V)	2, NP (1 V)	10
NP frame, verb second	7, NP (2, V)	8, V	8, V (2, NP)	10
NP-PP frame, verb final	9, NP	5, NP	5, NP	14
NP-PP frame, verb second	10, NP	8, NP (5, V)	10, NP (4, V)	14
Accuracy	66.2	63.2	66.2	
Coverage	95.7	97.05	100	
Correct Attachments	28.9 (71.1)%	20.9 (79.1)%	22.2 (77.8)%	

Table 4.10: Results on the experimental items development set for words and lemmas as input

To make these compounds more processable for the parser, they were reduced to their head words, which are more frequent. In most cases, this preserves the crucial semantic information. For example, *credit card* becomes *card*, which is also an acceptable term for such an item.

The DMM lemmatiser (Lorenz, 1996) which was used for lemmatisation also does compound analysis. Heuristically, the semantic head of a German compound noun is the rightmost component. These components are extracted and replace the compound nouns in the input items. In most cases, this works very accurately. However, there are a few examples in the development set that are not correctly decomposed by this heuristics, e.g. *Füllfederhalter*, *Füll-feder-halter* which correctly decomposes into Füll-Federhalter in the first instance, not into Füllfeder-Halter as DMM and the heuristics propose. The DMM morphology seems to have difficulty, too, to decide whether end *e*s and *n*s are case endings or part of the lemma, and in 12 out of 29 cases of words ending in *-e* or *-en* removes the last character. Quite frequently, this distorts word meaning, as for *Decke* – *Deck* (*blanket* – *deck*). The affected words are restored manually before they are processed by the semantic module to ensure the correct word meanings are passed to the semantic module.

Condition	Precision	Recall	F-Score	Tagging Acc.	Valid Sent.
Baseline	71.30	72.45	71.87	95.27	99.2
SC, add. data	67.11	68.28	67.68	89.08	99.5
SC, add. data, PT	74.39	77.26	75.79	-	92.8
SC, red. hds, add. counts	66.45	68.07	67.25	88.94	99.2
SC, red. hds, add. counts, PT	77.33	74.41	75.84	-	93.8

Table 4.11: Results on the NEGRA development set when compound nouns are reduced to their heads (red. hds)

For the lemmatised NEGRA development corpus, Table 4.11 shows that both F-Score and tagging accuracy rise by about three points when compound nouns are replaced by their heads. The upper bound for parser performance remains the same as for the version with full heads. This indicates that the reduction of compound nouns to the more frequent head nouns is a useful measure to tackle the sparse data problem and does not introduce additional errors.

Condition	Additional Data	Reduced Compounds	out of
NP-PP frame, verb final	3, NP	6, NP	10
NP-PP frame, verb second	8, NP (2, V)	8, NP (2 V)	10
NP frame, verb final	2, NP (1 V)	3, NP (2 V)	10
NP frame, verb second	8, V (2, NP)	8, V (2 NP)	10
NP-PP frame, verb final	5, NP	7, NP	14
NP-PP frame, verb second	10, NP (4, V)	10, NP (4, V)	14
Accuracy	66.2	76.5	
Coverage	100	100	
Correct Attachments	22.2 (77.8)%	21.2 (78.8)%	

Table 4.12: Results on the experimental items with reduced compounds

On the experimental items, accuracy rises to 76.5%, and 78.8% of attachments are correct assuming the reversed preferences in our data (see Table 4.12). This assumption is again justified by the clear pattern of attachment biases in the parser output. Only the verb final NP-subcategorisation condition does not show a reversal of attachment bias, but the attachment decisions are almost tied. Coverage remains at a perfect 100%.

4.6.3 Missing Grammar Rules

A third fundamental problem was found on inspection of the grammar rules. There are no rules to license PP-Attachment to the verb in the inverted word order conditions, e.g. in sentences like 3.4 above (*Neulich hörte ich, daß Iris die Rentnerin mit der Rockmusik störte.*). This sentence structure has simply not been seen in the NEGRA corpus. This is why the Baseline model almost always assigns NP-attachment and why the attachment figures are always worse for the verb final conditions than for the verb second conditions. The missing grammar rule is not a reason to disqualify the Baseline predictions, because they only mirror that one attachment alternative is so extremely scarce that it is not attested in the training data. Therefore, it is correct that the Baseline should predict the alternative attachment.

The underlying reason for the missing rule is that the grammar, being read off the corpus, does not generalise from names (NEs) to NPs, so there are separate rules for S $\rightarrow$ NP V and S $\rightarrow$ NE V. The latter rules are less frequent, so PP-Attachment to the verb is not accounted for at all in the verb final case. To solve this sparse data problem, a unary rule of the form NP $\rightarrow$ NE was introduced into the grammar. The frequency attached to the rule was extrapolated from the accumulated frequency of all grammar rules containing a name on the right hand side, except for rules that themselves license the formation of an NP.

When this new grammar rule is used with the subcategorised, lemmatised model, accuracy on the experimental items goes up to 92.6% (see Table 4.13). Comparison with the Baseline grammar shows the improvement of both accuracy and the number of correct attachment decisions that is brought about by the addition of subcategorisation information and the sparse data handling. The trend towards preferring the opposite attachment from what would be expected from the subcat preferences found in experimental data and used in Konieczny et al. (1997) becomes undeniable here. All verbs with a PP subcategorisation preference show a substantial preference for NP-

Condition	Baseline	NP $\rightarrow$ NE - Rule	out of
NP-PP frame, verb final	5, NP	10, NP	10
NP-PP frame, verb second	10, NP	6, NP (2 V)	10
NP frame, verb final	5, NP	7, V (2 NP)	10
NP frame, verb second	10, NP	8, V (2 NP)	10
NP-PP frame, verb final	12, NP	10, NP (2, V)	14
NP-PP, verb second	14, NP	11, NP (3, V)	14
Accuracy	82.4%	92.6%	
Coverage	100%	100%	
Correct Attachments	26.8 (73.2)%	17.5 (82.5)%	

Table 4.13: Results on experimental items with an NP $\rightarrow$ NE-rule

attachment, and all NP-subcategorisation verbs prefer verb attachment. Therefore, in the rest of the paper, the preferences as evident in our data will be used for our models, while the preferences assumed in Konieczny et al. (1997) will be applied for their results.

Since the NEGRA corpus does not contain the new rule, the augmented grammar cannot be tested on the NEGRA development set. The addition of the rule is purely a step to ensure correct coverage of the PP-Attachment cases.

With the manual addition of a generalising grammar rule, all three sparse data problems have been addressed.

4.7 Monitoring the Parsing Process

As a last step in building the syntactic model, it is necessary to extract the parser's current analyses at the PP, so the attachment preferences of the parser for verb-final clauses can be compared to the attachment preferences in the experimental data. For verb-second clauses, the results at the PP are identical to the final parses.

The parser is modified to return the chart (a record of the current state of the analysis) not only at the end of the sentence but also after the PP has been processed. All completed phrases and all incomplete analyses of sentences are returned so that the development of the two parses which account for the attachment alternatives can be monitored even when the analysis of the sentence is not complete.

In normal parsing, the one most probable structure would be extracted from the final chart and returned as the result. Here, we are interested in two output alternatives and their probabilities. Apart from these two structures, many more incorrect accounts for the sentence structure also coded in the parser's output. The task of expanding just the correct parses was done by a naive chart expansion algorithm.

The chart expansion algorithm extracts parses from the chart by recursively assembling their subphrases. Since the chart contains all active sentence edges at any given time, it soon becomes too big for exhaustive expansion and search. Therefore, the problem is reduced through two strategies:

- Only sentences generated by the rules that will lead to correct parses are expanded. This excludes the vast majority of sentence structures and brings the amount of analyses to be expanded down a more manageable level.

- Below the sentence level, there are no restrictions on the subphrases and therefore there is no limit on the number of parses for any subphrase. Every possible parse for the subphrases will be expanded, but only the n most probable ones are returned to the calling level of recursion. This ensures that space limitations are considered. Practically, n of five to ten has proven to be a good number to ensure that the correct analyses of both attachment alternatives are returned.

The more probable of the two expanded attachment structures is considered as parser's attachment choice. It is admissible to directly compare the probabilities of the alternative parses at the PP, even though they are for uncompleted structures. Since the set of possible completions is exactly the same for both alternatives when they are returned and since both will eventually be completed with exactly the same material, the sum of the probabilities of the potential completions is a constant over both attachment cases and can be neglected.

Extracting the verb final experimental items from Experiment 2 at first still caused problems because a number of different syntactic structures are used to arrive at constructions with embedded verb final sentences, which causes great ambiguity. In order to facilitate parsing and extraction, the syntactic structure of the top-level clauses from Experiment 2 was standardised. This of course does not affect the interesting regions of the sentences at all.

Table 4.14 shows that by actively choosing to look only at the correct parses, accuracy rises to 98.5% from the 92.8% that the final model reaches when only the best parses are considered. The reason for this is that for some of the sentences in the experimental items development set that were misparsed before, correct parses are now found. Not all of these parses are in line with the attachment bias of the condition, however, which is why the percentage of correct attachments goes down to 74.6%. This is a direct result of the increase in accuracy, which proves a mixed blessing.

Condition	Baseline	Best Correct Parses	out of
NP-PP frame, verb final	5, NP	8, NP (2 V)	10
NP-PP frame, verb second	10, NP	7, NP (3, V)	10
NP frame, verb final	5, NP	8, V (2 NP)	10
NP frame, verb second	10, NP	8, V (2 NP)	10
NP-PP frame, verb final	12, NP	10, NP (4, V)	14
NP-PP frame, verb second	14, NP	9, NP (4, V)	14
Accuracy	82.4%	98.5%	
Coverage	100%	100%	
Correct Attachments	26.8 (73.2)%	25.4 (74.6)%	

Table 4.14: Results on experimental items when both structural alternatives are extracted

4.8 Summary

The final syntactic module incorporates subcategorisation information through special verb tags. Sparse data problems are tackled by adding additional frame counts, reducing rare noun compounds to their heads and adding a new grammar rule that allows generalisation to existing rules in the cases where grammar rules are missing. The module is fairly reliable on the experimental items test set. It assigns the desired structures to 98.5% of the experimental items development set and predicts 74.5% of attachments correctly.

Performance on NEGRA has been traded for performance on the experimental items throughout. Experimentation with Perfect Tags shows, however, that especially the introduction of subcategorisation information would potentially lead to an improvement over the Baseline for the NEGRA data if enough training material were available.

The model now uses a grammar which incorporates subcategorisation information. Its input data is lemmatised and compound nouns have been reduced to their heads. An additional grammar rule has been introduced to ensure that all rules necessary to build correct parses for the experimental items exist.

As the performance of the module improves, it becomes more and more clear that the verb subcategorisation preferences from Konieczny et al. (1997) are indeed reversed in our data. From Chapter 5 on, the subcategorisation preferences will therefore be labelled as they appear in our data.

Chapter 5

Module II: Semantic Disambiguation

The parser accounts for the syntactic aspects of attachment quite reliably and with good coverage. Although the effects of semantic disambiguation can be inferred from the semantic bias of each condition, a semantic parser module that infers them from the materials themselves makes for a much more complete model. The second step in the development of the model is now to build a module that also accounts for the semantic disambiguation present in the experimental items. This module approximates the world knowledge that leads humans to make their final attachment decisions by looking at co-occurrence of words in a large corpus of language data. This makes it a *shallow* system, as it does not do a full semantic evaluation of the data. By using frequency information, the module keeps to the general spirit of the syntactic module, although we do not test the same strong claims about the influence of frequency on semantics that we test for syntax. The module's input is the two competing sentence structures that the parser outputs. Its task is to decide the final attachment of the PP on the basis of the constellation of the noun in the PP and the two attachment sites. Five different methods were evaluated for this task. They are described in detail in Section 5.2.

As a source for the frequency data needed by the measures, the NEGRA corpus was ruled out. Many of the the nouns in the experimental items are rare and 46% of the verbs are not accounted for in the NEGRA corpus at all, so it is improbable that co-occurrence counts from this corpus will be useful to make the right attachment decision. A bigger corpus is needed, and is provided in this case by the World Wide

Web. Even though the results for attachment disambiguation with Web counts are far worse than the results for counts from annotated data (see Section 2.3), this is the best option in the absence of an annotated corpus which covers the vocabulary in the experimental items.

In the next section, a pretest is described which shows that the performance of the semantic module does not decrease when only the heads of compound nouns are used. In Section 5.2, we introduce the measures used to make attachment decisions. Section 5.3 describes the strategies needed to deal with a potentially large number of queries due to the comparatively rich German morphology. Section 5.4 justifies the choice of `www.google.com` as search engine and the language restrictions imposed on the search engine.

Attachment decisions must be made for two cases: Verb-second sentences in which both the potential attachment sites for the PP (the verb and the noun) have been read in when the PP is processed, and verb-final sentences, where only the noun has been read and the verb is unseen yet.

Section 5.5 summarises the results for the different measures in the standard case with two known attachment sites. Section 5.6 describes how the measures were adapted to deal with the situation in German verb final sentences where only one of the attachment sites has been seen when a preliminary attachment decision has to be made.

5.1 Reduced Compound Nouns

The final version of the syntactic module uses input materials in which all compound nouns have been reduced to their heads. To exclude that this negatively affects the performance of the semantic module, the heads-only and full word conditions were compared using one of the measures (Mutual Information, see below). There was no difference between the full compounds and the heads-only condition when only lemmas were queried, but an improvement of the heads-only condition over the full compounds condition by two percentage points was observed when all morphological forms of the words were queried (see Section 5.3 on the form of the queries and the use of morphological information). Volk (2001) uses the same head reduction strategy for PP-attachment disambiguation as a sparse-data handling method, also with good results.

5.2 Measures

Most work in PP-Attachment to date uses similar configurational measures as the one introduced by Hindle and Rooth (1991). Generally speaking, the measures count whole attachment configurations and decide for the more frequent configuration. Optimally, these methods exploit structural annotation in corpora to determine attachment configurations. Since Web pages are not syntactically annotated, configuration counts have to rely on the adjacency of words in the documents. We also evaluate a new approach which relies on a more semantically inspired method that makes attachment decisions not on the basis of previously seen configurations but on the basis of semantic association between the attachment sites. This association is estimated from word co-occurrence counts in a large corpus. This approach seems suited to the task at hand because the items are meant to be clearly disambiguated by the semantic bias. It can therefore be hoped that one of the attachment sites, being semantically unrelated to the noun in the PP, will co-occur with that noun much less often than the other attachment site.

Five different methods of deciding the attachment with frequency counts are evaluated. For the standard case with two known attachment sites, these measures can be used unaltered (see Section 5.5). The strategies used to cope with the situation where only one attachment site is known are described in Section 5.6. Evaluation of the methods was done on the development set of experimental items. This allows an estimation of their performance on the test set without using that set before the final evaluation.

The first three measures are configurationally oriented, and the last two semantically oriented. Table 5.1 gives the formulae that are set into a ratio by the configurational measures. Table 5.2 shows the same for the semantically oriented measures. For each of these measures, the attachment is made according to which term in the ratio is greater: A larger term for attachment to the noun corresponds to NP-Attachment, a larger term for attachment to the verb to verb attachment. The measures are listed below:

- The Lexical Association Score is the ratio of the conditional probabilities of the preposition given the noun or the verb respectively (the *site*). This measure serves as a syntax-oriented baseline and captures the preference of the attachment site in question to be modified by a PP beginning with the preposition in question.

- Model 1 from Volk (2001) is the ratio of the frequency for the trigram *Noun, Preposition, Noun in PP* over the trigram *Verb, Preposition, Noun in PP*. This measure looks at the raw trigram co-occurrence frequencies to decide attachment.

- Model 2 from Volk (2001) is the same as Model 1 above, but normalises by the frequency of the attachment site. This takes into account that high-frequency attachment sites are more likely to co-occur with PPs in the first place.

- Pointwise Mutual Information (MI) is a measure from Information Theory. It is the the joint probability of the attachment site and the noun in the PP normalised by the prior probabilities of both the attachment site and the attachment candidate. The corpus size N is a constant here and is included to give the standard definition of the measure. MI measures how much information about one of the items is gained when the other is seen. This measure is used to approximate semantic knowledge rather than to count previously syntactic attachment sites. It has been used for the related problem of identifying collocations (words that appear together more often than chance, Church and Hanks (1990))

- Combined Conditional Probabilities (CCP) is the product of the conditional probabilities of the attachment site and the noun in the PP. It is quite similar to Mutual Information, and intuitively captures the fact that an attachment should be very probable if the joint probability of the words is high (i.e. the words often appear together) even when their co-occurrence is normalised by their frequencies. By squaring the joint probability term, it gives it more weight than MI.

Lex. Assoc.	Volk 1	Volk 2
$\frac{f(site,p)}{f(site)}$	$f(site,p,noun_{PP})$	$\frac{f(site,p,noun_{PP})}{f(site)}$

Table 5.1: Configurational measures to decide attachment decisions – $site$: attachment site (NP or verb), p: preposition, $noun_{PP}$: head noun of the PP

Mutual Information	Combined Conditional Probability
$log_2(\frac{f(site,noun_{PP})*N}{f(site)*f(noun_{PP})})$	$\frac{f(site,noun_{PP})}{f(site)} * \frac{f(site,noun_{PP})}{f(noun_{PP})}$

Table 5.2: Semantic measures to decide attachment decisions – $site$: attachment site (NP or verb), p: preposition, $noun_{PP}$: head noun of the PP, N: Corpus size

5.3 Assembling Web Queries

The word frequencies used by the measures are approximated by the number of documents on the Web that contain the word or words in question. German verbs and nouns have a comparatively rich morphology. Just querying the word forms that appear in the test data may lead to sparse or distorted frequency counts because only one form out of several possible ones has been considered. It is therefore desirable to query all forms of a word or at least all the most frequent ones. This is a form of *query expansion*, i.e. adding search terms to improve performance for a query.

A list of morphological forms for each word in the experimental item corpus was compiled with the MMorph generation facility, which is currently under construction (see Russell and Petitpierre (1995) on MMorph in general).

For the co-occurrence based semantic measures, this list of word forms can be used directly to query for the co-occurrence of any form of one word with any form of the other. For the configurational measures, the situation is different. The most accurate frequency estimates for configurations in the absence of syntactic annotation can be reached by querying for strings of words, which means that they have to appear in the retrieved documents in exactly the specified order. This has several drawbacks, however. For one, it does not allow for intervening modifiers. In German, different word orders also have to be explicitly accounted for. Secondly, every permutation of morphological form has to be spelt out, so the string has to be reformulated for singular and plural forms of different cases. This makes it problematic to query for a large number of different word forms. Additionally, the queries have to be of a form that actually occurs in the corpus.

As an example of combinatorial explosion incurred by naively built string queries, let us consider the formulation of the PP for configurational measures. The PP should have the form of *P Det N*, which opens a choice of definite, indefinite or no article in the singular and plural for the determiner alone. German determiners agree with the noun in case and gender. There are together 12 forms of the definite and indefinite determiner, so together with the no determiner option, there are 13 different determiners to be considered. Naively adding every form of the noun to every form of determiner adds up to 52 variations of the PP for a noun with four morphological forms. For an average of 10 forms per verb there are 520 naively expanded queries for verb attachment, with an additional 208 for NP attachment (again assuming four forms of the noun). This means about 700 queries per decision. Since Google restricts the number of queries to 1000 a day per user, this is not a feasible way of finding co-occurrence terms for larger numbers of items.

There are three ways of cutting down the number of queries. The one most obvious from the above example is to build noun phrases intelligently by paying attention to case and number agreement between the preposition, the determiner and the noun. This was done wherever PPs had to be formed. Fortunately, the preposition used in the experimental items, *mit,* constrains the following noun to the dative case, so only one form of the noun per singular and plural has to be queried for (the archaic dative forms ending in *-e* are infrequent and can be neglected). Additionally, gender information is used to find the correct forms of the definite and indefinite determiner. This pares the total number of alternatives for the PP from 52 down to five.

Another way to restrain the number of queries is to reduce the amount of word forms to be queried. This brings down the number of forms for the attachment sites. The third is not to query for strings, but to approximate string counts. This allows alternate word and phrase forms to be combined in one query and addresses the problem of intervening modifiers. These two strategies are described in detail in the next sections.

5.3.1 Reducing the Number of Word Forms per Query

As an example for word form reduction, Volk (2001) queries only for the word form encountered in the attachment task and its lemma. A manual check of the words in the set of experimental items revealed that the lemma is indeed one of the most frequent forms of German verbs and nouns, but not necessarily the only frequent or the most frequent form. For verbs, apart from the infinitive form, all the present forms are quite frequent, as are the third person forms in the simple past. Some of these forms double as past participle, which is used in the formation of other time forms. For nouns, the singular and plural nominative forms seem to be most frequent, as they also cover other cases (depending on the declension class). Explicitly marked genitive and dative forms tend to be less frequent.

It is desirable to query for at least all of the frequent forms to ensure that the results are representative. A second version of the list of morphological forms was therefore constructed which contains only the most frequent word forms for each entry. Now, the average number of entries per verb is about four, which already reduces the number of string queries per attachment site in the above example by more than half. This is a very welcome reduction in comparison with the full number of forms, while it still offers counts that rely on more than two spot checks of the distribution of the word and the other search terms.

In the following, performance with the original and reduced list of morphological items is tested for two example measures: Mutual Information for the semantically oriented measures, and the Volk 2 measure for the configurational measures.

For the configurational measures, performance improves on the reduced morphology set over the full set (see Table 5.3). This is because ineligible word forms such as inflected participles have been removed together with forms that are ambiguous. These word forms inflate the normalisation counts for the attachment sites while they cannot match the attachment configuration that has been queried for. For these measures, only the reduced set of morphological forms is therefore used to keep down the necessary number of queries as motivated above.

For the semantic measures, the full set of morphological forms can be used, because the queries are not for strings but for word co-occurrence in documents and extensive use of the Boolean operator *OR* can be made. This operator causes the search engine to look for documents matching any of the search terms, which reduces the number of queries. For example, to find all the co-occurrences of forms of the words *Rentnerin (pensioner)* and *Rockmusik (rock music)*, the query can be formulated as *(Rentnerin OR Rentnerinnen) AND Rockmusik*. Performance deteriorates when the reduced set of only the most frequent forms is used with the MI measure (see Table 5.3), probably because the inflected participle forms that are usually used as adjectives have been deleted. They are of importance for the semantic measures, because they capture word co-occurrence and not so much syntactic usage. The table also shows that querying for the wrong word forms can be worse than querying for just the word form from the input (MI, One Form versus Short morphology list).

Measure	Volk 2		Mutual Information		
Condition	Short	Full	One Form	Short	Full
Total correct	43	39	42	38	43
Percent correct	64.2%	58.2%	62.7%	56.7%	64.2%

Table 5.3: Results for Volk 2 and MI (two known attachment sites) with different morphology settings (only the input word (One Word), most frequent forms (Short), full list (Full))

5.3.2 Approximating String Queries

The second way of reducing the number of queries is not to query for strings of Site-Preposition-Noun configurations directly. This also tackles the problem of inflexible search terms that do not allow intervening modifiers or inverted word order.

Volk (2001) uses the NEAR operator available for AltaVista, which limits the distance between the query terms to 10 words. It does not restrict the ordering of the query terms, however, so that the resulting figures are a very rough approximation of the co-occurrence of Site, Preposition and PP head noun in the desired configuration.

(5.1)	*(...,*	*daß*	*Iris)*	*“die*	*Rentnerin*	*mit*	*dem*	*Rock*	*stört”*
	(...,	that	Iris)	the	pensioner	with	the	skirt	annoys

(5.2)	*“Rentnerin mit”*	+	*Rock*
	“pensioner with”	+	skirt

(5.3)	*Rentnerin*	+	*“mit dem Rock”*
	pensioner	+	“with the skirt”

(5.4)	*(Rentnerin*	*OR*	*Rentnerinnen=*	*AND*	*(“mit dem Rock”*	*OR*	*“mit Rock”*
	(pensioner	OR	pensioners)	AND	(“with the skirt”	OR	“with skirt”
	OR	*“mit einem Rock”*	*OR*	*“mit den Röcken”*	*OR*	*“mit Röcken”)*	
	OR	“with a skirt”	OR	“with the skirts”	OR	“with skirts”)	

Figure 5.1: Original, approximated and expanded queries

Here, the trigram counts are approximated by looking for “Site Preposition”+ Noun and Site+ “Preposition (Det) Noun”. Figure 5.1 gives an example of the original and approximated queries. Sentence 5.1 is the example attachment to be decided. Examples for query strings are *Rentnerin mit dem Rock* and *mit dem Rock stört*. Items 5.2 and 5.3 are approximations of the query strings for NP-Attachment, *die Rentnerin mit dem Rock*. Item 5.4 shows the query in 5.3 expanded with morphological forms. Splitting up the string query allows for the use of the OR operator on both subterms of the query. The query in 5.4 still has to be sent in two parts because Google allows only ten words (without Boolean terms) per query. This is very little, however, compared to the 208 queries that would be incurred by naive, fully combinatorial querying.

It appears that there is hardly any overlap between the two approximation terms from a few spot checks by hand. The overlaps would of course be exact matches for the trigram strings to be approximated. This means that the counts can be added without overestimating too badly. It also shows that even when using the Web as a corpus, there is still a sparse data problem. For the example case *“die Rentnerin mit* *dem Rock stören"*, neither string query (*"Rentnerin mit dem Rock"* and *"mit dem Rock stören"*) returns any counts. Splitting the strings into configurations as introduced above matches inverted sentences, too, and allows intervening modifiers, which helps to overcome the problem. Of course, the parts of the split strings can appear anywhere in the document, so there is no guarantee that the split strings actually stem from the same sentence. This makes the results an approximation.

5.4 Search Engines and Language Restriction

Measure	Mutual Information		Volk 2	
Condition	Google	AltaVista	Google	AltaVista
Total correct	43	35	43	37
Percent correct	64.2%	50.7%	64.2%	55.2%

Table 5.4: Results for MI and Volk Model 2 for different search engines

Frequency counts were collected from both `www.google.com` and `www.altavista.de`. Results for both semantic and configurational measures show that the counts from Google are more informative than counts from AltaVista. Table 5.4 summarises the results for the example measures MI and the Volk Model 2. Google probably searches more German pages than AltaVista, so its counts are less sparse. For English, there seem to be no big differences between the search engines (Lapata and Keller, 2003).

Measure	Mutual Information		Volk 2	
Condition	German	all	German	all
Total correct	46	43	41	43
Percent correct	68.7%	64.2%	61.1%	64.2%

Table 5.5: Results for two measures with and without restriction to German

Restricting the search to German data only results in an increase of performance for the MI measure, but for the Volk 2 measure, performance deteriorates (see Table

5.5). This is probably because the restricted morphology set used for this measure has already been controlled for homographs from other languages. Also, it appears that for many words the unigram frequencies for the attachment sites become smaller when Google search is restricted to German, while the trigram frequencies are not affected. This reduction is not in scale for both sites, which in many cases causes the decision to go wrong. Following these results, the configurational models were run with unrestricted search and the semantic models with German data only.

5.5 Two Known Attachment Sites

The case with two known attachment sites is the standard task covered in the literature, so all five measures can be used directly and without alteration on the verb second sentences from the experimental items development set.

As expected, the Web counts show good coverage – there is one instance of sparse data (no counts for either attachment alternative) for the co-occurrence based measures and two for the more restrictedly querying configurational measures. These instances are caused by a rare complex noun that the lemmatiser could not split. Even the Web corpus is not big enough in this case to furnish counts for an attachment that humans understand without problems. Using the reduced compound nouns that the syntactic module works with appears to have been a good strategy to avoid more serious sparse data problems for the Web counts.

In the evaluation, in cases where the semantic module cannot make a decision, the syntactic module's decision is accepted as default. Alternatively, it could be assumed that NP-attachment should be the default, but it seems more convincing to accept the syntactic preference if the semantic module is not able to make a decision at all. For all cases with valid outputs, the attachment alternative that received the numerically larger value becomes the semantic modules' decision.

One item that is biased towards verb attachment cannot be parsed correctly, so the default of always attaching to the NP is correct in slightly more than half of the cases. This makes the Baseline for the semantic module 50.7% correct attachments. The results for all five measures are listed in Table 5.6. Mutual Information, one of the semantic measures, perform best. The second semantic measure, Combined Conditional Probabilities, follows closely. The Lexical Association measure expectedly performs badly, only just outperforming the 50% Baseline. Volk Model 1 is on one level with CCP, and Model 2 does slightly worse. This is surprising, since it outperforms Model 1 in Volk (2001) and Lapata and Keller (2003). Compared to state-of-the art attachment disambiguation for German which is 73% (Volk, 2001) these values are still rather low.

For some measures, there appear to be tendencies to attach to one site rather than to another. Usually, the difference between NP and verb biased conditions is only by one or two correct attachments, so any interpretation has to be cautious. The Volk 2 measure shows better (or equal) performance in verb bias conditions than in NP bias conditions. The Lexical Association measure generally performs better in NP bias conditions than in verb bias conditions (and quite markedly so on the data for Experiment 2). For the other measures, no clear preferences are visible.

Since there are several measures which perform similarly well, we did a χ^2-test on the numbers of correct and incorrect decisions for each measure to further analyse their performance and to decide which measures to run on the test set. Comparisons were made both with the values for the best-performing measure and with the 50% Baseline. The χ^2 values and levels of significance are listed in Table 5.7.

Comparison with the Baseline shows that MI, the best-performing measure, is doing significantly better than the Baseline. For all other measures, the null hypothesis (the distribution of correct and incorrect decisions is the same as for the Baseline) cannot be rejected.

Comparisons of the three configurational measures and CCP with MI shows that there are no significant differences in the performance of all five measures.

Numerically, the Lexical Association measure is clearly the worst-performing measure. Its χ^2 values also show it to be closest to the Baseline and farthest away from the best-performing measures. It will therefore not be run on the test set. All other measures will also be tested on that set.

Condition	CCP	$I(w_1, w_2)$	Lex. Assoc.	Volk 1	Volk 2	out of
NP-PP frame, fin, V	2	5	0	1	4	5
NP-PP frame, fin, NP	3	3	4	3	2	5
NP-PP frame, 2nd, V	3	3	1	2	4	5
NP-PP frame, 2nd, NP	4	4	4	4	3	5
NP frame, fin, V	4	5	3	4	4	5
NP frame, fin, NP	2	2	4	3	4	5
NP frame, 2nd, V	4	5	2	4	5	5
NP frame, 2nd, NP	2	1	2	3	2	5
verb final, V	5	4	2	5	4	6
verb final, NP	5	5	7	6	4	7
verb second, V	5	5	1	4	4	7
verb second, NP	5	4	7	5	3	7
Total correct	44	46	37	44	43	67
Baseline	50.7%	50.7%	50.7%	50.7%	50.7%	
Correct Attachments	65.7%	68.7%	55.2%	65.7%	64.2%	

Table 5.6: Results for all five measures on the development set (Absolute number per condition and overall percentages of correct attachments)

	Best	50% Baseline
MI	–	3.75, $p \leq 0.05$
CCP	0.03, $p \leq 0.89$	2.48, $p \leq 0.11$
LA	1.12, $p \leq 0.29$	0.12, $p \leq 0.73$
Volk 1	0.03, $p \leq 0.89$	2.48, $p \leq 0.11$
Volk 2	0.13, $p \leq 0.72$	1.95, $p \leq 0.16$

Table 5.7: Values for χ^2 and levels of significance (df=1) for the five measures in comparison to the best-performing measure (MI) and the Baseline (50% correct) on the development set

5.6 One Known Attachment Site

The situation in German head-final clauses is more difficult than the standard case: When the PP is read, only one of the possible attachment sites, namely the noun, has been encountered, but it is quite clear that there will be another possible attachment site at the unseen verb. Konieczny et al. (1997) found processing difficulty in these cases when the PP was an implausible modifier of the noun, so it is obvious that immediate semantic evaluation sets in and has to be accounted for.

The problem at hand is to estimate the plausibility of the noun in the PP modifying the NP as opposed to modifying an as yet unseen verb. One way of estimating the probability of co-occurrence of the noun in the PP with any given verb is to average over the results for the noun in the PP and every possible verb to arrive at a "generic value" for verb attachment. It is obviously impossible to compute this value for every verb of German, so we restrict ourselves to just the verbs in the test and development set. This backoff was realised for all four models.

Another possibility, which is open only to the Combined Conditional Probability measure, is to use the prior probability of the noun of the PP as an estimate of its conditional probability with every possible verb. This probability can be used instead of the value for verb attachment to decide the attachment. This form of backoff is only applicable for the CCP measure, because the co-occurrence to be estimated is much more complex for the configurational models and the estimation method of MI does not arrive at values of comparable size to the prior. The prior probability for the head noun of the PP is computed as

$$P(noun_{PP}) = \frac{f(noun_{PP})}{N}$$

by dividing its frequency ($f(noun_{PP})$) by the size of the corpus (N), which is the total number of German documents searched. Since this number is not stated by Google, it was empirically established as the total number of pages searched divided by 100.

Table 5.8 gives the results for testing on the development set of items from Experiment 1. The items from Experiment 2 were not tested because the averaging procedure is extremely costly in terms of web queries. The performance of all measures is relatively uniform. The Combined Conditional Probability with simple backoff to the

	Prior	Average			
Condition	CCP	CCP	MI	Volk 1	Volk 2
NP-PP frame, V	3	4	2	0	3
NP-PP frame, NP	3	0	1	3	1
NP frame, V	4	5	4	4	5
NP frame, NP	2	2	1	2	2
Total correct	12	11	8	9	11
Baseline	50%	50%	50%	50%	50%
Correct Attachments	60%	55%	40%	45%	55%

Table 5.8: Results for different backoff procedures and different measures on the development set

prior shows the best results at 60% correct attachments. The CCP measure that uses averaging backoff and the Volk Model 2 with the same strategy share second place with just one correct decision less. The measure using the average Mutual Information of the noun in the PP and all verbs in the test and development sets does worst (40%), while the Volk 1 measure is slightly better.

Another set of χ^2 tests was done on these results just as for the two-site case. The χ^2 values and significance levels for comparison with the 50% Baseline and the best measure are shown in Table 5.9. None of the comparisons reaches significance. No measure significantly outperforms the 50% Baseline, and the difference in performance between measures is not significant, either. It cannot be expected that any of the measures will perform better on the test set, but the semantic module has to make a decision for attachments in verb final sentences. The best-performing measure on the training set, CCP with backoff to the prior probability of the noun in the PP, will therefore be used for modelling incremental attachment. All other measures will also be run on the test set, as their performance is not significantly different from the CCP measure's. This will establish whether the relatively good performance of CCP with backoff to the prior is a true trend.

	Best	Baseline
CCP, Prior	–	0.1, $p \leq 0.75$
CCP, Average	0.0, $p \leq 1$	0.0, $p \leq 1$
MI	0.90, $p \leq 0.34$	0.1, $p \leq 0.75$
V1	0.40, $p \leq 0.53$	0.0, $p \leq 1$
V2	0.0, $p \leq 1$	0.0, $p \leq 1$

Table 5.9: Values of χ^2 and levels of significance (df=1)for the five measures compared to the best measure and the Baseline on the development set (Exp. 1)

Chapter 6

Results and Discussion

This chapter presents the results for the model on the experimental items test set. The results for the purely syntactic module and for the semantic module are given separately, followed by the results for the model as a whole. To evaluate the predictions of the syntactic module, its attachment decisions are compared to the correct outcome that is given by the semantic attachment bias introduced by the noun in the PP. Where the the parser's attachment decision is not the same as the one required by semantic bias, processing difficulty is predicted.

The semantic module is then evaluated with regard to the percentage of attachments that it predicts correctly for every condition, that is how well it does in discovering the true biases.

Finally, both parts are brought together and the predictions of the full model are evaluated against the experimental data.

6.1 Syntactic Module – Results

On the NEGRA test set, the final syntactic model achieves an F-Score of 68.79 and coverage of 98.03%. The F-Score is better than the F = 67.25 on the development set, but coverage has gone down from 99.2%.

On the experimental items test set, all sentences are assigned some structure, so coverage is 100%. Two nouns from the test set are not in the lexicon with a noun meaning, so four sentences cannot be parsed correctly, though, which brings parser accuracy down to 95.5%.

All results for the parser are reported for the subcategorisation biases as they came out of the data used here, i.e. all NP-PP frame conditions correspond to Konieczny et al.'s NP frame conditions and vice versa (see Section 4.8).

6.1.1 Experiment 1

Table 6.1 summarises the results on the test set. It lists the parser's decisions at the PP for verb final and verb second sentences. This amounts to a decision in the absence of the verb for verb final sentences and at the end of the sentence for verb second sentences. We report the number of correct attachment decisions per condition and the overall percentage of correct decisions.

In verb final sentences, the parser always attaches the PP to the unseen verb. This is of course correct in 50% of all cases.

In verb second sentences, the picture is more diverse. For verbs with an NP subcategorisation preference, there is a bias towards NP attachment, so more attachments are predicted correctly in the NP bias condition than in the verb bias condition, both absolutely and in percentage points. However, the picture is not as clear as in the verb final sentences: 27% of attachments in the NP bias condition and 29% of attachments in the verb bias condition are to the verb. For verbs with a preference for PP objects, there is a also clear preference towards verb attachment, but 40 and 33% of attachments are to the NP. In sum, again 50% of attachments are correct, but this is not caused by the same categorical attachment preference as in the verb final condition.

6.1.2 Experiment 2

The verbs in this experiments are verbs with a preference for NP and PP objects in Konieczny et al.'s dataset and mostly verbs with an NP object preference in the subcategorisation data used here.

Table 6.2 summarises the results. As was the case with the items from Experiment 1, the parser always attaches the PP to the unseen verb in verb final sentences.

	verb final	verb second	out of
NP frame, V bias	7 (100%)	2 (29%)	7
NP frame, NP bias	0	5 (83%)	6
NP-PP frame, V bias	5 (100%)	3 (60%)	5
NP-PP frame, NP bias	0	2 (33%)	6
Total correct	12	12	24
Percent correct	50%	50%	
Baseline	50%	50%	

Table 6.1: Syntactic module: Correct Attachment decisions at the PP for data from Experiment 1

The preference for NP attachment is marked in the verb second conditions, but far from all attachments have been made to the NP in the NP bias condition (the attachment bias to the NP is almost perfect in the verb bias condition). This results in only 33% correct attachments overall.

	verb final	verb second	out of
V bias	9 (100%)	1 (11%)	9
NP bias	0	5 (56%)	9
Total correct	9	6	18
Percent correct	50%	33%	
Baseline	50%	50%	

Table 6.2: Syntactic module: Correct Attachment decisions at the PP for data from Experiment 2 (Percentage of correct decisions per condition in brackets)

6.2 Syntactic Module – Discussion

The results from Konieczny et al. (1997) that have been introduced in Chapter 3 can be summed up as follows:

- Verb final conditions: PPs with a semantic attachment bias towards attachment to the verb take longer to read than NP-biased PPs, so the initial preference is to attach a PP to the NP.
- Verb second conditions: Lexical preference effects show. After Konieczny et al.'s NP-PP frame verbs, NP-biased PPs take longer to read and vice versa for NP frame verbs. This means that the PP is initially attached in accordance with the verb's preference.

These are the effects that the model should reproduce. Generalising over the experiments, the syntax-based module predicts increased reading times (through conflicting attachment bias and outcome) for NP-biased PPs in verb-final sentences. It also predicts increased reading times whenever a semantic attachment bias does not match syntactic bias, i.e. a lexical preference effect exists.

The categorical attachment bias to the verb in verb-final sentences is exactly the opposite of the experimental results. The prediction of a lexical preference effect matches the findings from Konieczny et al. (1997) well.

These results can be visualised by the comparison of graphs for the results from Konieczny et al. (1997) and from the model. The Konieczny et al. figures are mean total Regression Path Durations in milliseconds, the figures for the model are the percentage of parser decisions that conflict with the semantic bias of the condition. Recall that these are assumed to cause difficulty, so a large amount of conflicting decisions predicts longer reading times.

Figure 6.1 contrasts the reading times and parser errors for verb final sentences of Experiment 1. The verb subcategorisation preferences in the graphs are per study, that is, the data from Konieczny et al. is labelled with their preferences, while the data for the model is labelled with the preferences found here. For the Konieczny et al. data, there is a significant increase in reading times for sentences with a verb-attachment bias. The data from the model shows exactly the opposite, as observed above.

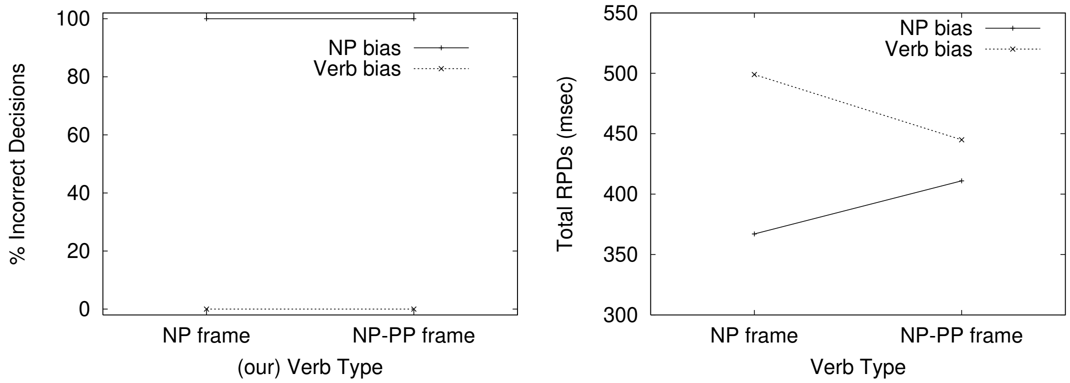

Figure 6.1: Experiment 1, verb final sentences: Error rates for the Syntactic Module (left) and mean reading times from Konieczny et al. (1997)

Figure 6.2 demonstrates the good fit of the model's prediction for verb second sentences in Experiment 2. The model shows the same interaction of subcategorisation preference and semantic bias as the data from Konieczny et al. (the effect is significant only by subjects in their data, though).

In Experiment 2, there was no variation of verb subcategorisation, so the reading times for verb final and verb second sentences are plotted together. The graphs for Experiment 2 (Fig. 6.3) show a replication in principle of the lexical preference effect. The significant effect of verb subcategorisation and semantic bias in the Konieczny et al. data is mirrored by the error rates for the parser, although it goes in the opposite direction because the verb bias is reversed. Konieczny et al. found an NP-PP preference for the verbs, which is why reading times are much higher for NP biased PPs. In our data, the verbs appear mainly biased towards a single NP object, so longer reading times are predicted for verb biased PPs. Again, strikingly, the model's prediction for the verb final case is wide off the mark.

6.2.1 Explanation of the Syntactic Module's Behaviour

Both the correct replication and the conflicting results can be explained by the probabilistic nature of the model and the characteristics of the underlying training data.

At the PP of verb-final sentences, the model stipulates attachment to the verb be-

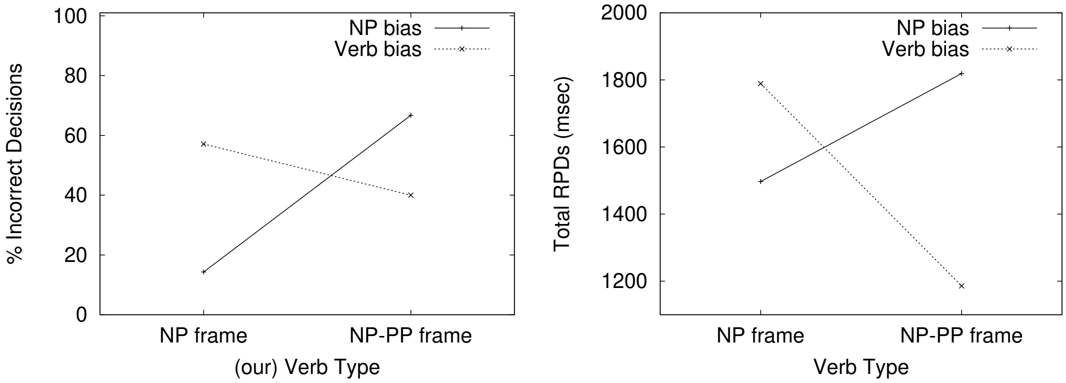

Figure 6.2: Experiment 1, verb second sentences: Error rates for the Syntactic Module (left) and mean reading times from Konieczny et al. (1997)

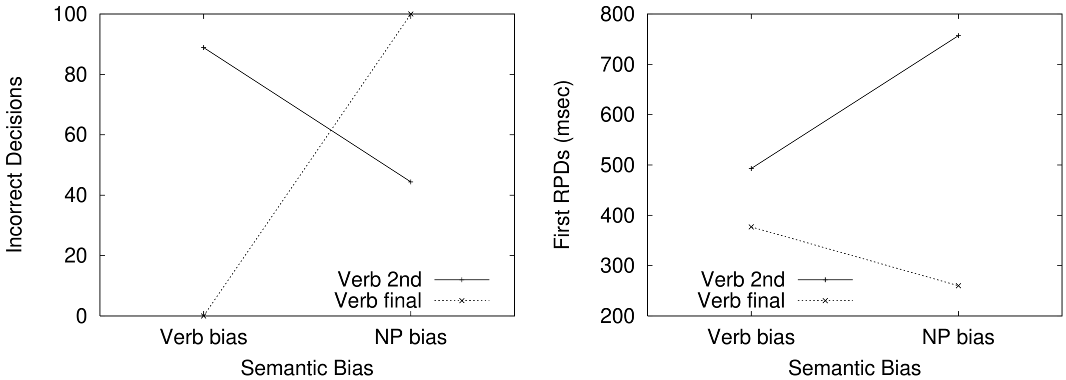

Figure 6.3: Experiment 2, verb second and verb final sentences: Error rates for the Syntactic Module (left) and mean reading times from Konieczny et al. (1997)

cause the probability of seeing an NP and a PP separately (i.e. a verb-attached PP) is higher than seeing them as a complex NP. This is caused by two factors. For one, there is an extra rule involved in creating the complex NP that does not have to be taken into account in postulating verb attachment. Recall that NEGRA trees are flat and do not stipulate a VP node in most sentences. Since the probability of a parse is the product of the rule probabilities involved, a structure with an extra rule is always less probable than one with fewer rules.

However, the problem is not purely one of the annotation scheme. Recall that there are twice as many instances of PP attachment to the verb as to the NP in NEGRA, both absolutely and in percentages. This means that in the data, attachment to the verb is actually more probable than attachment to the NP. Together, these biases drown out the overall preference for sentences without verb attachment that biases attachment towards a complex NP. (The sentence rule without PP as a sister of the verb is three times more probable than the one with a PP modifying the verb.)

In both verb final and verb second sentences, the verb's preference for one of the subcategorisation frames decides the *final* outcome of the syntactic module's attachment decision for the whole sentence. The only difference is that in verb second sentences, the verb's influence has already been taken into account when the PP is read and can influence the *initial* attachment decision. Verbs with a high preference for seeing just an NP as their object make the verb attachment structure so improbable that even the preference for seeing the NP and the PP as separate phrases and attaching the PP to the verb cannot switch the ranking. The attachment of the PP therefore is made to the NP. If the verb has a subcategorisation preference for the NP-PP frame, attachment to the verb becomes overwhelmingly more probable and the correct decision is made. The imperfect attachment decisions in verb second sentences can therefore be explained by a minority of verbs with an attachment preference that does not match the general preference in their condition. Also, some of the verbs do not show a marked preference at all, but are equibiased between one and two objects. For these verbs, the balance between a general preference against verb attachment and the higher probability of a simple NP and PP is sensitive and can be tipped in one way or another by small differences in frame probability.

6.2.2 Implications for Statistical Models

From these results, there are three conclusions with immediate relevance to statistical models. Firstly, they present a renewed caveat regarding the equivalence of preference data from corpora and completion studies. Although a reliable correlation between these sources of data has been shown for English (BNC, Lapata et al. (2001)), this result apparently does not hold for all corpora. Both the NEGRA corpus and the corpus used for the extraction of the subcategorisation lexicon show an outright reversal of subcategorisation preferences with regard to preferences from completion studies. As argued in 4.3.1 above, a probable reason for this inconsistency is that the BNC is a balanced corpus which contains samples of a wide variety of written and spoken language from different genres and situations. It therefore approximates every-day language usage better than a corpus consisting of newspaper text only, such as the corpora directly and indirectly used here. Since statistical models are only as good as the data they are based on, our results underline the necessity of basing models on reliable language data. For German, however, there is no balanced corpus, and the situation is similar for many other languages, which makes accurate statistical modelling difficult.

Secondly, our results highlight the importance of the choice of modelling framework. In our case, there is an interference between the annotation scheme and the PCFG framework of the model. The flat annotation scheme creates an unbalanced amount of phrase structure rules involved in forming the two attachment alternatives and the PCFG used as a model reacts very sensitively to imbalances of this kind. It is probably impossible to avoid such imbalances for all psycholinguistically interesting phenomena, whatever the annotation scheme, so the sole use of a PCFG as the backbone of modelling tools is called into question. Approaches such as Sturt et al. (2003) or McRae et al. (1998) which focus on decisions about attachment to unranked structures avoid problems of the kind encountered here.

Thirdly, there is another problem with PCFGs as a modelling framework. In order to correctly model the attachment preference to the NP in verb final sentences, our model would have needed corpus data with a substantial bias to NP-attachment. The NEGRA corpus, however, does not show this preference. This may be atypical, as Volk (2001) reports a bias towards NP-attachment in his German test corpus. However, it demonstrates that purely statistical models rely on production data showing the same attachment preferences as initial parsing decisions. However, not all initial attachment decisions can be modelled by late attachment preferences. People seem to initially attach newly read NPs as direct objects to preceding verbs whatever those verbs' subcategorisation preferences are (Mitchell, 1987; Pickering et al., 2000). The PP-Attachment preference to the NP in German verb final clauses can be seen as a related phenomenon. Note that this problem is not related to the reliability of corpus frequencies in any way – the problem is caused by a difference between initial and late attachment preferences.

This phenomenon can be interpreted in two ways: Either, it is an instance of Tuning, where people attach the NP according to overall structural preferences (transitive verbs are overall most frequent) as indicated Sturt et al. (2003), or it is caused by a general parsing principle. Two motivations for such a principle have been put forward. Konieczny et al. (1994) motivate this immediate attachment by the possibility of immediate semantic interpretation of the complete input, Pickering et al. (2000) trace it to a minimisation of costly forms of reanalysis, stating that the parser chooses the analysis that is most easily falsifiable, i.e. most *informative*. Since immediate semantic processing can reveal clues to the correctness or incorrectness of the current analysis, these two theories are not necessarily in conflict. If the phenomenon is indeed caused by a general parsing principle, the fit of initial decision and global structural preference might be coincidental and might possibly not hold in all cases. This would be starting point for a strategy of discerning between the two explanations.

Whatever the explanation for the immediate attachment phenomenon, it cannot be modelled by a PCFG. PCFGs only respond to global structural preferences for object attachment if they are not lexicalised and have no notion of verb subcategorisation information. Information of this kind is necessary, however, to model phenomena like PP-Attachment, where they influence the attachment process. The necessity of taking global and verb-specific structural preferences into account at different points in the timecourse of processing a single input word casts an interesting light on the Grain Size Problem: Different grain sizes of structural preference information seem to be used for initial and final attachment decisions during the processing of a single new input word.

If the immediate attachment phenomenon is interpreted to be due to some general parsing strategy, again PCFGs have no mechanism to take such a strategy into account. Indeed, it would have to be explicitly modelled by any probabilistic model.

The initial attachment decisions have been modelled successfully by Crocker and Brants (2000) for the NP/S-ambiguity, but the authors state explicitly that this was the case because their model, relying on a PCFG backbone, favoured the syntactically simpler alternative that involved fewer rule applications. For our model, this inbuilt bias was harmful rather than beneficial, so it should not be seen as a generally admissible bias towards simpler structure.

The last two points call into question the use of PCFGs to model the human sentence processor. Probabilistic models have to pay attention to lexical and global preferences alike to model the initial attachment preference outlined above or implement general parsing strategies like the preference for immediate attachment. It has become clear above that the PCFG-based models introduced in Chapter 2 (Jurafsky, 1996; Brants and Crocker, 2000; Hale, 2001) cannot model the initial attachment effects. For the models by Narayanan and Jurafsky (2001) and McRae et al. (1998), this depends on the choice and weighting of constraints. The model by Sturt et al. (2003) that allows the decision-making network to choose its own attachment criteria has been shown to correctly model the initial preference for the direct object reading in the NP/S-ambiguity, because the network's attachment decision apparently is influenced by global as well as lexical attachment preferences.

6.3 Semantic Module – Results

Below are the detailed results for the behaviour of the semantic module with one and two known attachment sites. For the case of only one known attachment site, all five measures were run again on the test set because their performance was not significantly different on the development set. For the case of two known attachment sites, the four best-performing models, namely MI, the combined conditional probability measure and the Volk Models 1 and 2 were run.

6.3.1 Two Known Attachment Sites

For the two attachment site case, there is a shift in ranks for the four models. Table 6.3 shows the performance of the measures by condition. On the development set, MI had been best numerically and had even significantly outperformed the Baseline. The two Volk models performed almost equally well, along with the CCP measure, but worse than MI. On the test set, the picture is quite different. The Volk 2 measure does much better than any of the other measures, while the Volk 1 measure, that had been performing equally well on the development set, is now on a par with Mutual Information and CCP. The preferences of attachment have also changed, so they are probably just due to chance variation. On the test set, the CCP measure seems to have a tendency to attach to the NP that it did not show on the development set, while the Volk 1 measure shows a tendency towards verb attachment. For the MI and Volk 2 measures it is hard to make out a tendency, although the Volk 2 measure showed a relatively clear preference towards verb attachment on the development set.

Again, a series of χ^2 tests was done on the results for the measures. Table 6.4 shows the outcome of the tests for the measures' results on the test set. Even though the Volk 1 measure clearly outperforms the others numerically, it does not significantly outperform the Baseline. It is the only measure, however, that in any way approaches significance at $p \leq 0.09$. Also, the gap in performance between the Volk 1 measure and the others could still be attributable to variance.

6.3.2 One Known Attachment Site

For verb-final sentences, where there is only one known attachment site when the PP is read, the combined conditional probabilities measure and the prior probability do as well on the test set as on the development set. Table 6.5 shows the performance per condition for all verb-final sentences.

On the development set, the CCP measure with backoff to the prior probability of the head noun of the PP had performed best numerically, closely followed by CCP with backoff to an average value for attachment to the verb and by the Volk 2 measure with the same backoff strategy. There was no significant difference between the five

Condition	CCP	MI	Volk 1	Volk 2	out of
NP frame, fin, V	2	2	2	3	7
NP frame, fin, NP	4	4	5	1	6
NP frame, 2nd, V	1	5	2	3	7
NP frame, 2nd, NP	2	3	4	4	6
NP-PP frame, fin, V	2	3	2	5	5
NP-PP frame, fin, NP	4	1	3	6	6
NP-PP frame, 2nd, V	2	3	2	4	5
NP-PP frame, 2nd, NP	3	2	3	3	6
verb final, V	4	6	3	6	9
verb final, NP	8	6	7	6	9
verb 2nd, V	5	6	4	6	9
verb 2nd, NP	8	4	8	7	9
Total correct	45	45	45	54	84
Baseline	50 %	50 %	50 %	50 %	
Percentage correct	53.6%	53.6%	53.6%	64.3%	

Table 6.3: Semantic Module, Two Known Attachment Sites: Results on the Test Set (frame preferences for our data)

	Best	Baseline
MI	1.57, $p \leq 0.21$	0.1, $p \leq 0.75$
CCP	1.57, $p \leq 0.21$	0.1, $p \leq 0.75$
Volk 1	1.57, $p \leq 0.21$	0.1, $p \leq 0.75$
Volk 2	–	2.94, $p \leq 0.09$

Table 6.4: Two Known Attachment Sites: Values of χ^2 and levels of significance (df=1) for the comparison to the best measure and the Baseline on the test set

	Prior	Average				
Condition	CPP	CCP	MI	Volk 1	Volk 2	out of
Exp 1, NP frame, V	3	4	3	3	5	7
Exp 1, NP frame, NP	6	0	4	4	5	6
Exp 1, NP-PP frame, V	2	3	3	3	5	5
Exp 1, NP-PP frame, NP	4	2	3	4	1	6
Exp 2, V	4	7	2	1	5	9
Exp 2, NP	5	4	7	8	4	9
Total correct	24	20	22	23	25	42
Baseline	50%	50%	50%	50%	50%	
Percentage correct	57.1%	47.6%	52.4%	54.8%	59.5%	

Table 6.5: Semantic Module, One Known Attachment Site: Results for the five measures (with different backoff procedures) on test data (frame preferences for our data)

measures, and none significantly outperformed the Baseline. The measures were rerun here to generate prediction data for the one attachment site case and to see whether there is any apparent trend of performance.

The best-performing measure on the test set is the Volk 2 measure, closely followed by the CCP/prior measure. The Volk 1 measure is again doing slightly better here than MI. The most drastic change in performance is the drop of the CCP/Average model to the level of worst performance from comparatively good performance on the development set.

A series of χ^2 tests agains the Baseline and the best performing measure not surprisingly shows no significant differences in the performance of the measures. Also, no measure outperforms the Baseline of 50% correct attachments. The χ^2 values and levels of significance are given in Table 6.6

	Best	Baseline
CCP, Prior	0.00, $p \leq 1.00$	0.19, $p \leq 0.66$
CCP, Average	0.77, $p \leq 0.38$	0.00, $p \leq 1.00$
MI	0.19, $p \leq 0.66$	0.00, $p \leq 1.00$
Volk 1	0.05, $p \leq 0.82$	0.05, $p \leq 0.83$
Volk 2	–	0.43, $p \leq 0.51$

Table 6.6: One Known Attachment Site: Values of χ^2 and levels of significance (df=1) for the comparison to the best measure and the Baseline on the test set

6.4 Semantic Module – Discussion

6.4.1 Two Known Attachment Sites

There was no significant difference in the performance of the five different measures on either the development or test set. Numerically, the Mutual Information measure performed better on the development set. It also was the only measure that significantly outperformed the Baseline on either set. The Volk 2 measure performed numerically best on the test set and also performed most consistently over both sets.

A probable reason for the disappointing performance of Mutual Information (and the related CCP measure) on the test set is that the semantic measures ignore syntactic information that the configurational measures preserve. For example, *Kind (child)* and *Märchen (fairy tale)* tend to co-occur in many documents, whereas *Kind* modified by *Märchen* is very rare. The large co-occurrence figure causes a high Mutual Information value, while the configurational measures take the much lower number of cases into account in which *Kind* is actually modified by *Märchen*. This allows the configurational models to take into account that sometimes, attachment to one of the sites is inacceptable not due to lack of semantic association, but because the instrument or modifier role introduced by the preposition is implausible. Semantic implausibility of attachment on the other hand is usually mirrored well in co-occurrence counts of the attachment sites and the noun in the PP and is therefore recognised well by the semantic measures. In the development set, there seem to have been more cases where attachment is decided on grounds of semantic implausibility than in the test set.

The Volk 1 model is the second model to show great inconsistency of performance on the test and development set. It also profited from the larger number of cases in the development set where one of the attachment sites very rarely co-occurs with the noun in the PP, thus making raw co-occurrence counts reliable predictors of attachment. Also, many of the attachments in the development set seem to have been to the more frequent attachment site. In this case as well, the raw co-occurrence frequencies are reliable enough as a basis for decision. Normalising by the frequency of the attachment site for each count does no harm, as the results for the Volk 2 model show, but it is not indispensable. On the test set, there are fewer such cases and normalising by the frequency of the attachment site is truly beneficial. Note that the semantic measures normalise by both the frequency of the attachment site and by the frequency of the noun in the PP. Their problems are not caused by taking the counts at face value, but by the fact that semantic association as apparent in Web counts is not always a good predictor of attachment.

There is no clear manifestation of an attachment preference for the semantic measures. Both Volk measures show a preference towards verb attachment once, but since the number of trials is quite small overall, this trend has to be regarded cautiously.

In sum, the Volk 2 measure, while failing to significantly beat the Baseline, shows the most constant performance of all measures. It is not affected by the differences between the test and development set as the other measures are. The semantic measures fail to demonstrate superior performance over the configuration based measures, which is mostly due to their reliance on semantic association, which appears to be not always a good predictor of attachment decisions.

6.4.2 One Known Attachment Site

On the test set, the relatively good performance of the CCP measure with backoff to the prior and of the Volk 2 measure are confirmed, as well as the bad performance of MI on the development set. The CCP measure with backoff to an average value for verb attachment shows a drop in performance on the test set as the semantic measures do for the standard case with two attachment sites. The poor performance of all models with regard to the baseline is also confirmed.

For the CCP model, which was tested with two types of backoff to extrapolate the strength of the attachment preference to the verb, the backoff to the prior probability of the noun in the PP seems to yield more constant performance than the backoff to an average verb attachment value. This is probably caused in part by the general difficulty the semantic measures have on the test set, and in part by the particular sample of verbs used for the backoff. They are in all probability not representative for German verbs with an NP- or PP-subcategorisation bias.The general poor performance of the other models also calls into question the usefulness of the averaging backoff, which probably also fails because of the unrepresentative sample of verbs that is being averaged over.

A third possible approach to the disambiguation of attachment decisions when only one attachment site is known would be to determine a threshold for the value for NP attachment and to attach the PP to the verb if that threshold is not reached. This would eliminate the necessity of averaging over a probably skewed sample of verbs. For the semantic measures, this approach is not applicable, however, because the Mutual Information values (and similarly the values for CCP) vary among sentences. Therefore, a low Mutual Information value for NP-Attachment on its own does not allow a definite attachment decision. The threshold approach was tried on the Volk 2 measure, with discouraging results. A threshold that accounts for 75% correct attachments on the development set (0.001) did not seem to be appropriate for the test set at all. Only 43% of attachments were correct that were made according to this threshold. A variation of this approach might be to set the threshold dynamically for every PP by seeing how strongly the noun in the PP is prefers attachment to the NP in comparison with other attachment sites. For this, we would need to average over a number of attachment sites for the threshold. Of course, this is costly in terms of web queries, which is why it could not be tested here. Also, it again poses the problem of finding a good sample of nouns for comparison to the original attachment site.

6.5 Predictions of the Full Model

In this section, the syntax-based decisions of the syntactic module and the decisions of the shallow semantic module are combined. Because of the categorical verb attachment bias of the syntactic module, the preference for attaching the PP to the NP in verb final sentences cannot be modelled correctly by the full model, either. The lexical preference effect that arises when verb subcategorisation preferences and semantic attachment bias clash is modelled correctly by the full model.

The figures to be compared with Konieczny et al.'s data here are the percentages of correct decisions by the semantic module that are in conflict with the parser's decision. In the case of conflict between a correct semantic module decision and the parser output, longer reading times are predicted, as above. Note that not all correct decisions by the semantic module necessarily conflict with the parser's output for verb second sentences, even if syntactic and semantic bias of the condition are not the same. This is because the parser does not assign the same attachment to every sentence in a condition, as detailed in Section 6.2.1 above. For this reason, it is necessary to take the percentage of correct decisions into account that actually override the parser's decision. This also means that correct predictions of reading time data that the full model makes are not trivial, because not every correct decision necessarily corrects the parser's output.

Verb final sentences In verb final sentences, the parser's preference is always attachment to the verb, so correct decisions for verb attachment always concur with the syntactic module, and correct decisions for NP-attachment always conflict with it. Predictions for the attachment decisions in the verb final case of only one known attachment site are approximated by the best-performing measure on the development set, namely the Combined Conditional Probability with backoff to the prior probability of the noun in the PP.

Table 6.7 gives the numbers of corrected parser decisions and the percentage these numbers make up of the total correct decisions per condition. These values confirm the predictions of the syntactic module with perfect semantic disambiguation. For Experiment 1 and 2 alike, the full model predicts longer reading times for an attachment of the PP to the NP.

Condition	CPP, Prior
Exp 1, NP frame, V	0%
Exp 1, NP frame, NP	100% (6)
Exp 1, NP-PP frame, V	0%
Exp 1, NP-PP frame, NP	100% (4)
Exp 2, V	0%
Exp 2, NP	100% (5)

Table 6.7: Semantic Module, Verb Final Sentences: Number of corrected parser decisions and percentage of all correct decisions they make up

Figure 6.4 allows a graphical comparison of the reading time predictions of the combined syntactic and semantic module and the experimental data for the verb final sentences of Experiment 1. As in the evaluation of the predictions of the syntactic module above, the complete model's predictions are contradicted by the data from Konieczny et al. (1997).

Verb second sentences The correct modelling of the lexical preference effect is shown exemplary for both the Mutual Information measure and for the Volk 2 measure. These are the best-performing measures on the development and test set respectively.

Table 6.8 gives the amount of semantic module decisions that corrected decisions by the syntactic module and the percentage these make up of the total number of correct module decisions for the condition. For both the Volk 2 and the Mutual Information measure, more parser decisions are corrected in the verb bias condition of NP-subcategorising verbs than in the NP bias condition. This points to longer reading times for verb attached PPs. Likewise, more parser decisions were corrected in the NP bias condition of the NP-PP-subcategorising verbs than in the verb bias condition. Here, a semantic disambiguation towards NP-attachment is predicted to lead to problems. In the data for the NP-frame verbs from Experiment 2, this pattern is clearly repeated.

Condition	Volk 2	MI
Exp 1: NP frame, V	67% (2)	80% (4)
Exp 1: NP frame, NP	0% (0)	33% (1)
Exp 1: NP-PP frame, V	50% (1)	67% (2)
Exp 1: NP-PP frame, NP	33% (1)	50% (1)
Exp 2: V bias	100% (6)	83% (5)
Exp 2: NP bias	43% (3)	50% (2)

Table 6.8: Semantic Module, Verb Second Sentences: Number of corrected parser decisions and percentage of all correct decisions those make up

Figure 6.5 contrasts the graphs for the predictions of the Volk 2 and Mutual Information measure with the graph for the original figures from Konieczny et al.'s data. As above, the subcategorisation preferences are labelled as they appear in each source of data. The two alternative semantic modules correctly predict the interaction between subcategorisation preference and semantic bias that shows in the Konieczny et al. data.

For the sake of completeness, Figure 6.6 contrasts the predictions of the full model with Volk 2 and MI as semantic modules with the actual reading time data found by Konieczny et al. Again, there is a replication of the lexical preference effect only in principle, because of the reversed verb subcategorisation preferences. The CCP/prior measure is used in both cases for the prediction of attachments in verb final sentences. As was to be expected from Table 6.7 and from the discussion of the syntactic module's general behaviour, the NP attachment preference in verb final sentences again cannot be modelled.

In sum, the complete model makes the same predictions as the syntactic module alone, accounting correctly for subcategorisation preferences in verb second sentences and not accounting for the attachment preference to the NP in verb final sentences. This result is reached even though the alternative measures for the semantic module do not significantly outperform the Baseline.

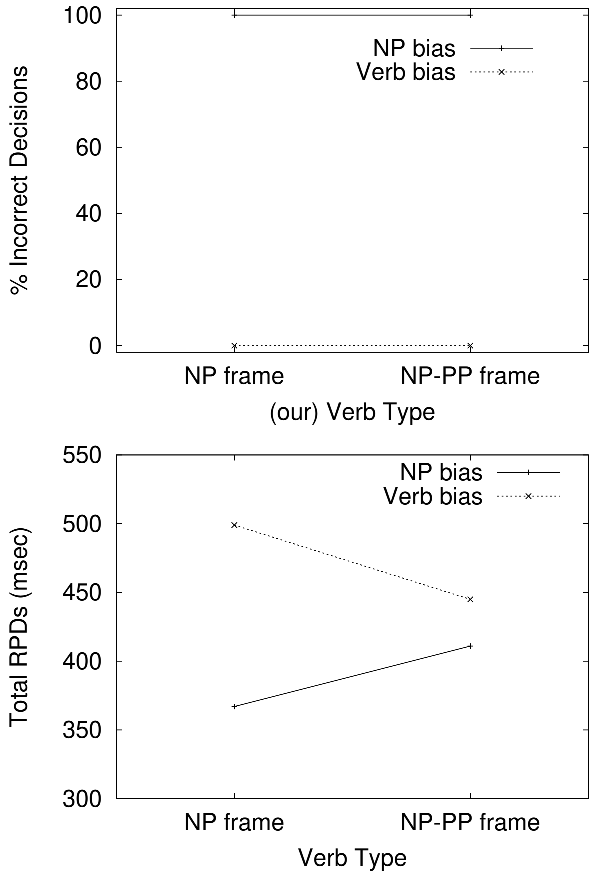

Figure 6.4: Experiment 1, verb final sentences: Predictions of the CCP/Prior model (top) in comparison with the Konieczny et al. (1997) data (bottom)

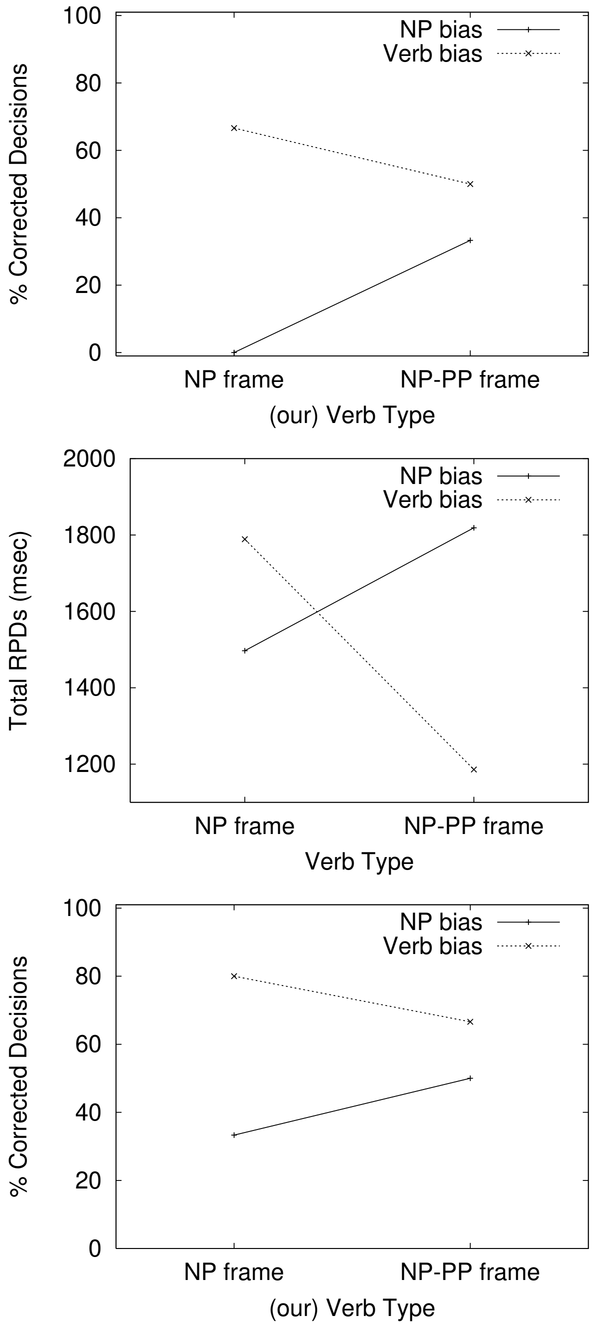

Figure 6.5: Experiment 1, verb second sentences: Predictions of the Volk 2 model (top) and MI (bottom) in comparison with the Konieczny et al. (1997) data (middle)

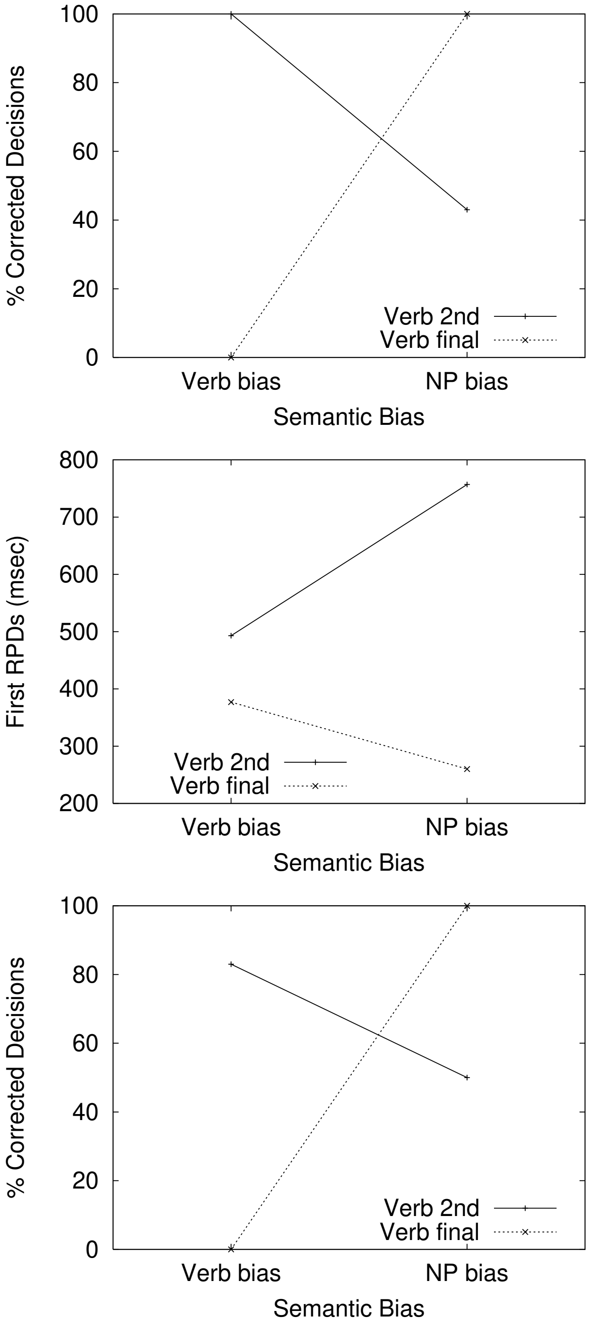

Figure 6.6: Experiment 2: Predictions of the full model in comparison with the Konieczny et al. (1997) data (middle) – verb second sentences: Volk 2 (top), MI (bottom), verb final sentences: CCP/Prior (top and bottom)

Chapter 7

Conclusions

This Thesis has described a two-stage probabilistic model of human incremental attachment decisions in German PP-Attachment. German verb final sentences offer an opportunity to study initial PP-Attachment decisions in the absence of the sentence head, which also constitutes the second possible attachment site. In this situation, the PP is preferentially attached to the (existing) NP site by human readers (Konieczny and Hemforth, 2000).

The first module of our model is a purely syntactic and based on a PCFG parser, which guarantees wide coverage of language data. The second module uses shallow semantics and attempts to determine the semantically correct attachment of the PP. Two candidate strategies of deciding PP-Attachment were evaluated: One works on the basis of previously seen instances of the PP and the attachment sites (configurational approach), the other relies on differing semantic closeness of the head noun of the PP and the attachment sites (semantic approach). The configurational model proved more reliable than the model assessing semantic closeness. This is probably because raw semantic association is a less good predictor of attachment preferences than the plausibility of the noun in the PP to be an instrument (of the verb site) or modifier (of the noun site), which is better approximated by the configurational models. While both types of model fail to significantly outperform the chance baseline on the test set, their predictions with regard to the number of parser decisions that are revised by semantic bias still allow successful evaluation of the full model.

The first stage of the model was evaluated on its own and with the semantic module.

Both versions of the model correctly account for human attachment decisions in verb second sentences, while both fail to account for the initial preference to attach the PP to the seen NP in verb final sentences, where the verb is yet unseen when the PP is read.

The reasons for the partial failure to correctly model experimental data are innate in the present combination of modelling approach and data. The immediate cause of the wrong predictions lies in the flat annotation scheme of the corpus, which does not stipulate the same number of nodes for NP and verb attachment. The PCFG backbone of the model assigns a higher probability to the structural alternative with fewer rule applications, i.e. verb attachment (Section 6.2.1). Additionally, there is a general preference for verb attachment evident in the corpus data, which is not necessarily typical for German (Section 4.3.1). These two biases cause an initial preference for verb attachment in verb final sentences, where the verb's subcategorisation preference only comes into play later on.

The model's partial failure highlights a second problem with PCFGs as a modelling framework. PCFGs use attachment preferences of only one grain size (e.g. verb subcategorisation preferences in lexicalised PCFGs *or* global attachment preferences in unlexicalised PCFGs) to model initial attachment decisions. There are example phenomena of initial attachment decisions in the literature which cannot be modelled using only one of the grain sizes. For example, the human parser apparently initially ignores subcategorisation preferences, and always attaches newly-read NPs to the preceding verb (Mitchell, 1987; Pickering et al., 2000). The PP-Attachment preference in verb final sentences can possibly be seen as another instance of an immediate attachment phenomenon.

This behaviour can be interpreted as initially following general structural preferences for the most common structure (i.e. a direct object reading as for transitive verbs) and switching to lexical preferences of the verb only slightly later. Alternatively, the human parser's eagerness to attach incoming heads to existing sites can be interpreted as falling out of a general parsing strategy such as *Parametrised Head Attachment* (Konieczny et al., 1994) or *Informativity* (Pickering et al., 2000) that allow immediate semantic evaluation (Parametrised Head Attachment) or quick falsification of the attachment decision (Informativity).

Since PCFGs cannot take global and lexicalised preferences into account in a way that would account for the immediate attachment preference and since they do not incorporate any general parsing principle, they are not good tools to model initial attachment phenomena like these.

Additionally, in our data we found a reversal of general and verb-specific attachment preferences as established through psycholinguistic studies (see Sections 4.3.1 and 4.8). This is a caveat for the claim that the results from completion and corpus studies are well correlated, at least for unbalanced corpora as the ones used here.

7.1 Future Work

Future work will mainly address the semantic module. Its current overall performance is unsatisfactory, even though its predictions in combination with the syntactic module model the experimental data well for verb second sentences (verb final sentences cannot be modelled accurately because the parser predicts the wrong initial attachment). The configurational measures have shown consistent but mediocre performance, while the semantic measures did well on the development set and badly on the test set. A strategy for general improvement of the semantic module could be to first optimise each class of measures separately and then to find a way of combining the best configurational and semantic measures if the configurational measures can still profit from the combination.

Configurational methods are much more accurate when their data can be induced from annotated corpora instead of being approximated by string adjacency in web documents. The main reason for our use of Web counts was that the existing annotated corpus does not cover the vocabulary of the experimental items (almost 50% of the verbs are not accounted for). This problem can be partly overcome by generalising to semantic noun classes. Instead of counting configurations of *pensioner* and *rock music*, one would count configurations of words falling into the classes *human* and *music*. Such classes are usually inferred from WordNet (Miller et al., 1990), a computer-readable semantic ontology for English. A similar resource exists for German (GermaNet, Hamp and Feldwig (1997)). In cases where a word is not covered by the ontology or where counts are still sparse, it is still possible to backoff to Web counts.

Another interesting way of determining plausible modifiers of nouns is not by co-occurrence frequency in corpora, but through a feature based approach. McRae et al. (1997) describe the collection of sets of feature descriptions of nouns from subjects. An example feature for *car* could be *has wheels*. The features could give an insight into typical attributes and modifiers of nouns, namely *wheels* as a good modifier for *car* in this case. This resource only exists for a limited amount of English nouns, though, and does not specify a probability distribution over features.

To determine the plausibility of the PP as an object of the verb, data from FrameNet (Baker et al., 1998) might be used to bolster sparse counts from NEGRA. A German corpus with FrameNet annotation is currently being developed (Erk et al., 2003). FrameNet specifies roles for objects, so PP objects which have a role in FrameNet can be extracted for each verb along with their frequencies. Then, the WordNet classes of their head nouns can be determined and lookup can be done to see whether any new combination of preposition and semantic class of the head word is a plausible filler of a role the verb specifies.

The semantic measures might also be improved by generalising over word classes because this relieves remaining sparse data problems. Semantic closeness can additionally be induced from relationships in WordNet. However, some sections of WordNet appear to be more fine-grained than others, so semantic closeness defined by the length of paths between words can be treacherous (Stetina and Nagao, 1997).

For cases of only one known attachment site, the improved combined measure with the threshold approach outlined in 6.4.2 is a starting point for improvement. A generally improved measure might in itself improve performance on the one-attachment site task, while the dynamic threshold might outperform a static threshold and even the averaging backoff.

The main weakness of the syntactic module, namely its wrong predictions for verb final sentences, can only be addressed by building a new model that is not based on the PCFG framework, as discussed in Section 6.2.2 and in the beginning of this chapter.

An additional weakness that can be addressed within the current framework the syntactic module's decrease in performance relative to the Baseline model. The Perfect Tag upper bound shows room for improvement over the Baseline, but a lack of training data, which leads to inaccurate tagging especially of verbs keeps the parser from reaching that optimal performance. The only solution for this problem is a larger training corpus or a smaller verb tag set. The verb tag set is already the smallest set that still meaningfully differentiates between subcategorisation frames. A larger training corpus is now available in form of the TIGER corpus (Brants et al., 2002). This is the successor of the NEGRA corpus which uses the same format of annotation but is already twice as large as NEGRA (40,000 sentences). There is no overlap between the corpora, so there are 60,000 sentences of consistently annotated German corpus data available now. Also, the annotation of grammatical functions in the TIGER corpus makes a distinction is made between PP complements and adjuncts, which makes verb frame induction and annotation much easier.

As a last point, the model has been built and tested with just one ambiguity phenomenon in mind. It would be interesting to test it on other ambiguities and extend it if necessary.

Appendix A

Experimental Items: Development and Test Set

This Appendix contains the development and test set of experimental items used in the experiments. The version given here contains no spillover region and no adjectives in the PP (see Section 4.4). Compound nouns have not been reduced to their heads (see Section 4.6.2). The original sentences appear in Konieczny et al. (1997).

A.1 Development Set – Experiment 1

NP-PP frame, verb final:

V bias

Neulich hörte ich, daß Thomas den Bauernschrank mit dem Pinsel bemalte.
Man erzählte mir, daß Marion die Torte mit der Spritztülle verzierte.
Gestern erfuhr ich, daß Karl das Heft mit dem Füllfederhalter beschriftete.
Mir wurde erzählt, daß Veronika die Tischdecke mit der Nadel bestickte.
Neulich hörte ich, daß Herbert die Tür mit dem Schrank versperrte.

NP bias

Gestern erfuhr ich, daß Iris die Rentnerin mit dem Rock störte.
Mir wurde erzählt, daß Hartmut das Mädchen mit dem Gesicht folterte.
Gestern erfuhr ich, daß Barbara das Reh mit dem Brandzeichen beobachtete.
Gestern erfuhr ich, daß Karl das Heft mit der Papierstärke beschriftete.

Mir wurde erzählt, daß Veronika die Tischdecke mit dem Rand bestickte.

NP-PP frame, verb second:

V bias

Iris störte die Rentnerin mit der Rockmusik.
Oliver bewarf die Wand mit dem Schneeball.
Anton belustigte das Publikum mit der Vorstellung.
Barbara beobachtete das Reh mit dem Fernglas.
Marion verzierte die Torte mit der Spritztülle.

NP bias

Oliver bewarf die Wand mit dem Fenster.
Anton belustigte das Publikum mit der Erwartung.
Barbara beobachtete das Reh mit dem Brandzeichen.
Karl beschriftete das Heft mit der Papierstärke.
Veronika bestickte die Tischdecke mit dem Rand.

NP frame, verb final:

V bias

Gestern erfuhr ich, daß Franz den Dackel mit der Krücke stieß.
Man sagte mir, daß Martin den Bankraub mit der Einsatztruppe untersuchte.
Man sagte mir, daß Heike die Natter mit dem Teleobjektiv erblickte.
Mir wurde erzählte, daß Sabine die Schatulle mit dem Weihnachtsgeld erwarb.
Gestern erfuhr ich, daß Nicole den Jungen mit dem Lied tröstete.

NP bias

Gestern erfuhr ich, daß Franz den Dackel mit dem Fell stieß.
Man sagte mir, daß Gabi den Pullover mit der Knopfleiste strickte.
Man sagte mir, daß Norbert das Haus mit dem Erker bewachte.
Neulich hörte ich, daß Hannah das Kind mit der Stupsnase ängstigte.
Mir wurde erzählte, daß Sabine die Schatulle mit dem Geheimfach erwarb.

NP-PP frame, verb second:

V bias

Franz stieß den Dackel mit der Krücke.
Gabi strickte den Pullover mit der Maschine.
Florian brachte den Blumenstrauß mit dem Wagen.
Sabine erwarb die Schatulle mit dem Weihnachtsgeld.
Nicole tröstete den Jungen mit dem Lied.

NP bias

Martin untersuchte den Bankraub mit dem Schaden.
Susanne beschenkte das Kind mit dem Zopf.
Gabi strickte den Pullover mit der Knopfleiste.
Heike erblickte die Natter mit dem Giftzahn.
Nicole tröstete den Jungen mit dem Bein.

A.2 Development Set – Experiment 2

NP-PP frame, verb final

V bias

Daß Bruno den Hasen mit dem Gewehr erschoß, war der Anfang vom Ende seiner Verbrecherlaufbahn.
Daß Felix das Brot mit dem Messer schnitt, wirkte etwas stilisiert.
Als Martin den Rahmen mit der Laubsäge bastelte, fiel ihm etwas ein.
Als Rita die Katze mit der Dreiklanghupe erschreckte, geriet sie beinahe in Panik.
Weil Robbi den Produzenten mit dem Gitarrensolo beeindruckte, kam ihm eine Idee.
Daß Tim den Spaziergänger mit dem Motorgeräusch belästigte, war so nicht geplant.
Daß Luise das Essen mit der Kreditkarte bezahlte, erregte Aufmerksamkeit.

NP bias

Daß Felix den Koch mit dem Messer knebelte, wirkte etwas stilisiert.
Weil Paul den Schüler mit der Zigarette verprügelte, merkte dieser zunächst nichts von dem Vorfall.
Daß Sarah die Lampe mit der Gasflamme löschte, kam ihr selbst etwas kitschig vor.
Weil Robbi das Musikstück mit dem Gitarrensolo bearbeitete, kam ihm eine Idee.

Daß Tim das Tonband mit dem Motorgeräusch zerstörte, war so nicht geplant.
Da Jochen den Eimer mit dem Wischwasser beklebte, bekam er später Ärger.
Während Richard die Wohnung mit der Fußbodenheizung versicherte, ärgerte er sich über die hohen Nebenkosten.

NP-PP frame, verb second

V bias

Peter ängstigte das Kind mit dem Schauermärchen.
Felix schnitt das Brot mit dem Messer.
Martin bastelte den Rahmen mit der Laubsäge.
Rita erschreckte die Katze mit der Dreiklanghupe.
Tim belästigte den Spaziergänger mit dem Motorgeräusch.
Richard regelte die Temperatur mit der Fußbodenheizung.
Sabine flog das Paket mit dem Privatjet.

NP bias

Felix knebelte den Koch mit dem Messer.
Paul verprügelte den Schüler mit der Zigarette.
Martin verpackte den Werkzeugkoffer mit der Laubsäge.
Sarah löschte die Lampe mit der Gasflamme.
Rita zerkratzte den Wagen mit der Dreiklanghupe.
Tim zerstörte das Tonband mit dem Motorgeräusch.
Jochen beklebte den Eimer mit dem Wischwasser.

A.3 Test Set – Experiment 1

NP-PP frame, verb final:

V bias

Mir wurde erzählt, daß Hartmut das Mädchen mit der Daumenschraube folterte.
Gestern erfuhr ich, daß Iris die Rentnerin mit der Rockmusik störte.
Man sagte mir, daß Anton das Publikum mit der Vorstellung belustigte.
Man sagte mir, daß Ingrid den Freund mit dem Anruf erfreute.
Man sagte mir, daß Oliver die Wand mit dem Schneeball bewarf.

Gestern erfuhr ich, daß Barbara das Reh mit dem Fernglas beobachtete.
Mir wurde erzählt, daß Ulla die Großmutter mit dem Kuß beglückte.

NP bias

Man sagte mir, daß Ingrid den Freund mit dem Schnupfen erfreute.
Neulich hörte ich, daß Herbert die Tür mit dem Abziehbild versperrte.
Man sagte mir, daß Oliver die Wand mit dem Fenster bewarf.
Man erzählte mir, daß Marion die Torte mit dem Mokkageschmack verzierte.
Man sagte mir, daß Anton das Publikum mit der Erwartung belustigte.
Neulich hörte ich, daß Thomas den Bauernschrank mit der Tür bemalte.
Mir wurde erzählt, daß Ulla die Großmutter mit der Gicht beglückte.

NP-PP frame, verb second:

V bias

Karl beschriftete das Heft mit dem Füllfederhalter.
Thomas bemalte den Bauernschrank mit dem Pinsel.
Herbert versperrte die Tür mit dem Schrank.
Ingrid erfreute den Freund mit dem Anruf.
Hartmut folterte das Mädchen mit der Daumenschraube.
Veronika bestickte die Tischdecke mit der Nadel.
Ulla beglückte die Großmutter mit dem Kuß.

NP bias

Iris störte die Rentnerin mit dem Rock.
Marion verzierte die Torte mit dem Mokkageschmack.
Ingrid erfreute den Freund mit dem Schnupfen.
Thomas bemalte den Bauernschrank mit der Tür.
Hartmut folterte das Mädchen mit dem Gesicht.
Herbert versperrte die Tür mit dem Abziehbild.
Ulla beglückte die Großmutter mit der Gicht.

NP frame, verb final:

V bias

Man sagte mir, daß Gabi den Pullover mit der Maschine strickte.
Man erzählte mir, daß Helmut den Patienten mit der Salbe verarztete.
Neulich hörte ich, daß Hannah das Kind mit dem Schauermärchen ängstigte.
Man sagte mir, daß Norbert das Haus mit dem Gewehr bewachte.

Gestern erfuhr ich, daß Florian den Blumenstrauß mit dem Wagen brachte.
Man sagte mir, daß Susanne das Kind mit dem Präsent beschenkte.

NP bias

Man sagte mir, daß Martin den Bankraub mit dem Schaden untersuchte.
Man sagte mir, daß Susanne das Kind mit dem Zopf beschenkte.
Gestern erfuhr ich, daß Florian den Blumenstrauß mit der Rose brachte.
Man erzählte mir, daß Helmut den Patienten mit der Wunde verarztete.
Man sagte mir, daß Heike die Natter mit dem Giftzahn erblickte.
Gestern erfuhr ich, daß Nicole den Jungen mit dem Bein tröstete.

NP frame, verb second:

V bias

Hannah ängstigte das Kind mit dem Schauermächen.
Norbert bewachte das Haus mit dem Gewehr.
Heike erblickte die Natter mit dem Teleobjektiv.
Susanne beschenkte das Kind mit dem Präsent.
Martin untersuchte den Bankraub mit der Einsatztruppe.
Helmut verarztete den Patienten mit der Salbe.

NP bias

Franz stieß den Dackel mit dem Fell.
Florian brachte den Blumenstrauß mit dem Wagen.
Hannah ängstigte das Kind mit der Stupsnase.
Sabine erwarb die Schatulle mit dem Geheimfach.
Helmut verarztete den Patienten mit der Wunde.
Norbert bewachte das Haus mit dem Erker.

A.4 Test Set – Experiment 2

NP-PP frame, verb final:

V bias

Man erzählte mir, daß Sabine das Paket mit dem Privatjet flog.
Man erzählte mir, daß Peter das Kind mit dem Schauermärchen ängstigte.
Man erzählte mir, daß Claudia den Scherbenhaufen mit den Breitreifen überrollte.
Man erzählte mir, daß Susanne die Grenze mit dem Hubschrauber erreichte.
Man erzählte mir, daß Jochen das Fenster mit dem Wischwasser bespritzte.

Man erzählte mir, daß Hans das Loch mit der Steinplatte bedeckte.
Man erzählte mir, daß Paul die Decke mit der Zigarette versengte.
Man erzählte mir, daß Richard die Temperatur mit der Fußbodenheizung regelte.
Man erzählte mir, daß Sarah das Papier mit der Gasflamme entzündete.

NP bias

Man erzählte mir, daß Claudia den Wagen mit den Breitreifen untersuchte.
Man erzählte mir, daß Sabine den Piloten mit dem Privatjet überprüfte.
Man erzählte mir, daß Luise die Brieftasche mit der Kreditkarte verdreckte.
Man erzählte mir, daß Peter das Buch mit dem Schauermärchen verstand.
Man erzählte mir, daß Bruno den Jäger mit dem Gewehr fesselte.
Man erzählte mir, daß Susanne die Plattform mit dem Hubschrauber sah.
Man erzählte mir, daß Rita den Wagen mit der Dreiklanghupe zerkratzte.
Man erzählte mir, daß Martin den Rahmen mit der Laubsäge bastelte.
Man erzählte mir, daß Hans das Loch mit der Steinplatte säuberte.

NP-PP frame, verb second:

V bias

Susanne erreichte die Grenze mit dem Hubschrauber.
Claudia überrollte den Scherbenhaufen mit den Breitreifen.
Robbi beeindruckte den Produzenten mit dem Gitarrensolo.
Jochen bespritzte das Fenster mit dem Wischwasser.
Paul versengte die Decke mit der Zigarette.
Luise bezahlte das Essen mit der Kreditkarte.
Bruno erschoß den Hasen mit dem Gewehr.
Hans bedeckte das Loch mit der Steinplatte.
Sarah entzündete das Papier mit der Gasflamme.

NP bias

Sabine überprüfte den Piloten mit dem Privatjet.
Luise verdreckte die Brieftasche mit der Kreditkarte.
Claudia untersuchte den Sportwagen mit den Breitreifen.
Richard versicherte die Wohnung mit der Fußbodenheizung.
Bruno fesselte den Jäger mit dem Gewehr.
Robbi bearbeitete das Musikstück mit dem Gitarrensolo.
Susanne sah die Plattform mit dem Hubschrauber.
Peter verstand das Buch mit dem Schauermärchen.
Hans säuberte das Loch mit der Steinplatte.

Bibliography

Gerry T. M. Altmann and Mark J. Steedman. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238, 1988.

John R. Anderson. Is human cognition adaptive? *Behavioural and Brain Sciences*, 14: 471–517, 1991.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, Montreal, Canada, 1998.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 2002.

Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, 2000.

Thorsten Brants and Matthew W. Crocker. Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th International Conference on Computational Linguistics*, 2000.

Eric Brill and Philip Resnik. A rule based approach to prepositional phrase attachment disambiguation. In *Proceedings of the Fifth International Conference on Computational Linguistics*, 1994.

Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of NAACL-2000*, 2000.

Nicholas Chater, Matthew Crocker, and Martin Pickering. The rational analysis of inquiry: The case for parsing. In Nicholas Chater and Michael Oaksford, editors, *Rational Models of Cognition*. Oxford University Press, 1998.

Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 1(16):22–20, 1990.

Michael Collins. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, 1997.

Michael Collins and James Brooks. Prepositional attachment through a backed-off model. In David Yarovsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 27–38, Somerset, New Jersey, 1995. Association for Computational Linguistics.

Matthew W. Crocker and Thorsten Brants. Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, 29(6):647–669, 2000.

Fernando Cuetos and Don C. Mitchell. Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, 30:73–105, 1988.

Amit Dubey and Frank Keller. Probabilistic parsing for German using sister-head dependencies. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.*, Sapporo, 2003.

Susan A. Duffy, Robin K. Morris, and Keith Rayner. Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, 27:429–446, 1988.

Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of ACL-03*, Sapporo, Japan, 2003.

Lynn Frazier and Keith Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14:178–210, 1982.

Susan M. Garnsey, Neal J. Pearlmutter, Elizabeth Myers, and Melanie A. Lotocky. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37:58–93, 1997.

Edward Gibson, Carson T. Schütze, and Ariel Salomon. The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research*, 25:59–92, 1996.

Herbert P. Grice. Logic and conversation. In Donald Davidson and Gilbert Harman, editors, *The Logic of Grammar*. Dickenson, 1975.

John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 159–166, 2001.

Birgit Hamp and Helmut Feldwig. GermaNet – A lexical-semantic net for German. In Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorick Wilks, editors, *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Association for Computational Linguistics, New Brunswick, New Jersey, 1997.

Donald Hindle and Mats Rooth. Structural ambiguity and lexical relations. In *Meeting of the Association for Computational Linguistics*, pages 229–236, 1991.

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

Lars Konieczny and Barbara Hemforth. Modifier attachment in German: Relative clauses and prepositional phrases. In A. Kennedy, R. Radach, D. Heller, and J. Pynte, editors, *Reading as a Perceptual Process*, pages 517–527. Elsevier, 2000.

Lars Konieczny, Barbara Hemforth, Christoph Scheepers, and Gerhard Strube. PP-Attachment in German: Results from eye movement studies. In J.M.Findlay, R. Walker, and R. W. Kentridge, editors, *Eye Movement Research. Mechanisms, processes, and applications*. Elsevier, 1995.

Lars Konieczny, Barbara Hemforth, Christoph Scheepers, and Gerhard Strube. The role of lexical heads in parsing: Evidence from German. *Language and Cognitive Processes*, 2/3(12):307–348, 1997.

Lars Konieczny, Christoph Scheepers, Barbara Hemforth, and Gerhard Strube. Semantikorientierte Syntaxverarbeitung. In S. Felix, C. Habel, and G. Rickheit, editors, *Kognitive Linguistik: Repräsentationen und Prozesse*. Westdeutscher Verlag, 1994.

Maria Lapata and Frank Keller. Evaluating the performance of unsupervised web-based models for a range of NLP tasks. Unpublished Manuscript, 2003.

Maria Lapata, Frank Keller, and Sabine Schulte im Walde. Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research*, 30(4):419–435, 2001.

Oliver Lorenz. Automatische Wortformerkennung für das Deutsche im Rahmen von Malaga. Master's thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 1996.

Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. Lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703, 1994.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, 1993.

Ken McRae, Virgina R. de Sa, and Mark S. Seidenberg. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2), 1997.

Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38:283–312, 1998.

Paola Merlo. A corpus-based analysis of verb continuation frequencies for syntatic processing. *Journal of Psycholinguistic Research*, 23(6):435–457, 1994.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 5 papers on WordNet. `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`, 1990.

Don C. Mitchell. Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart, editor, *Attention and perfomance XII*, pages 601–618. Erlbaum, 1987.

Don C. Mitchell, Fernando Cuetos, Martin M. Corley, and Marc Brysbaert. Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, 24, 1995.

Srini Narayanan and Daniel Jurafsky. A Bayesian model predicts parse preferences and reading times in sentence comprehension. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS2001)*, 2001.

Martin J. Pickering, Matthew J. Traxler, and Matthew W. Crocker. Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language*, 43:447–475, 2000.

Adwait Ratnaparkhi. Statistical models for unsupervised prepositional phrase attachment. In *Proceedings of the Seventeenth International Conference on Computational Linguistics*, Montreal, 1998.

Adwait Ratnaparkhi and Salim Roukos. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Workshop on Human Language Technology*, 1994.

Douglas Roland and Daniel Jurafsky. Verb sense and verb subcategorization probabilities. In Paola Merlo and Suzanne Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. John Benjamins, 2002.

Graham Russell and Dominique Petitpierre. MMORPH - The Multext Morphology Program. MULTEXT deliverable report for task 2.3.1, 1995.

Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS, 1995. URL `http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html`.

Helmut Schmid. Lopar - Design and Implementation, 2000. URL `http://www.ims.uni-stuttgart.de/~schmid/lopar.ps`.

Carson T. Schütze and Edward Gibson. Argumenthood and English prepositional phrase attachment. *Journal of Memory and Language*, 40:409–431, 1999.

Sabine Schulte im Walde. A subcategorisation lexicon for German verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, volume IV, pages 1351–1357, Las Palmas de Gran Canaria, Spain, 2002.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC, USA, 1997.

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

Michael Spivey-Knowlton. *Integration of visual and linguistic information: Human data and model simulations*. PhD thesis, University of Rochester, Rochester, N.Y., 1997.

Michael Spivey-Knowlton and Julie C. Sedivy. Resolving attachment ambiguities with multiple constraints. *Cognition*, 55:227–267, 1995.

Jiri Stetina and Makoto Nagao. Corpus based PP attachment ambiguity resolution with a semantic dictionary. In *Proceedings of the 5th Workshop on very large corpora*, pages 66–80, 1997.

Gerhard Strube, Barbara Hemforth, and Heike Wrobel. Resolution of structural ambiguities in sentence comprehension: Online analysis of syntactic, lexical, and semantic effects. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pages 558–565, 1989.

Patrick Sturt, Fabrizio Costa, Vincenzo Lombardo, and Paolo Frasconi. Learning first-pass structural attachment preferences with dynamic grammars and recursive neural nets. *Cognition*, 2003.

John C. Trueswell. The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, 35:566–585, 1996.

Martin Volk. Scaling up. Using the WWW to resolve PP attachment ambiguities. In *Proceedings of Konvens-2000*, Ilmenau, 2000.

Martin Volk. Exploiting the WWW as a corpus to resolve PP attachment ambiguities. In *Proceedings of Corpus Linguistics 2001*, Lancaster, 2001.