Universität des Saarlandes
Philosophische Fakultät II – Sprach-, Literatur- und Kulturwissenschaften
Fachrichtung 4.7 – Allgemeine Linguistik
Computerlinguistik
Postfach 15 11 50
66041 Saarbrücken

Diplomarbeit

Evaluating and Combining Approaches to Selectional Preference Acquisition

Carsten Brockmann

25. September 2002

Angefertigt unter Leitung von
Prof. Dr. Manfred Pinkal
und
Dr. Maria Lapata

Erklärung

Ich erkläre an Eides statt, dass ich die Diplomarbeit mit dem Titel „Evaluating and Combining Approaches to Selectional Preference Acquisition“ selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und alle den benutzten Quellen wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Saarbrücken, 25. September 2002

Carsten Brockmann

Acknowledgements

I would like to thank my professor Manfred Pinkal for his supervision and for having created with his team a great research environment which made this work possible. I am also deeply grateful to my supervisor Maria Lapata who helped me getting interested in the subject in the first place and gave me excellent advice whenever I needed it.

I would also like to thank Günter Neumann of the German Research Center for Artificial Intelligence (DFKI) GmbH for his support concerning the SMES system, and Garance Paris for her assistance in evaluating SMES. I am thankful to Frank Keller for his help regarding the web-based experiment, and to him and Maria Lapata for letting me use their implementation of Clark and Weir's algorithm as a basis for my own source code. Further thanks go to CLT Sprachtechnologie GmbH and its managing director Gerd Fliedner for granting access to the language corpora that were created at this company.

I am grateful to Sebastian Ritterbusch and Stefan Thater for reading through several drafts of this thesis and providing me with extensive and extremely helpful comments. Also, a big thank you goes to Christoph Scheepers for advice on a lot of issues.

I would like to thank my fellows and colleagues at Saarland University for anything: Bettina Braun, Ralph Debusmann, Philipp Detemple, Peter Dienes, Amit Dubey, Markus Egg, Katrin Erk, Frederik Fouvry, Malte and Ute Gabsdil, Kerstin Hadelich, Tilman Jäger, Kerstin Klöckner, Pia Knöferle, Alexander Koller, Andrea Kowalski, Geert-Jan M. Kruijff, Ivana Kruijff-Korbayova, Tronje Pannhausen, Cristian Pietsch, C.J. Rupp, Joachim Sauer, Werner Saurer, Julia Singer, Kristina Striegnitz, and Olga Uryupina. This list is not supposed to be exhaustive.

Finally, thanks to my family for all their support. Without them, of course, I would not even be here, let alone be able to write a thesis about selectional preference acquisition.

Summary

Selectional preferences are graded constraints that predicates impose upon their arguments. They can be used for tasks like parse ranking or word sense disambiguation, or can be included in dictionary entries. In the last decade, various methods have been suggested to automatically acquire selectional preference information from corpora.

There have been many evaluations of the automatically acquired selectional preference information. However, most have been conducted for English, only for the verb/direct object relation, and looking at algorithms in isolation. Open questions are whether the methods are valid cross-linguistically, how they perform on other grammatical relations, and how they perform in combination. This information can justify or rebut the usefulness of state of the art techniques for selectional preference acquisition.

The research presented in this thesis attempts to fill these gaps. I conducted a magnitude estimation experiment over the world-wide web to elicit human judgments on the acceptability of 90 German sentences with intransitive and transitive verbs, and verbs subcategorizing for a PP object. These judgments were compared to the preferences computed by five approaches to automatic selectional preference acquisition: Two frequency-based measures, selectional association (Resnik 1993), tree cut models (Li and Abe 1998), and the similarity-class measure (Clark and Weir 2001).

I found that there exist significant linear correlations between the human judgments and different single algorithms, depending on the grammatical relation being observed. Furthermore, when methods are combined by a principal components factor analysis and multiple regression, they outperform the results of a single approach when judging all of the test sentences.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Selectional Restrictions versus Selectional Preferences

Selectional restrictions are constraints that a predicate imposes on the arguments it can take. They have first been introduced in the work of Katz and Fodor (1963) and Chomsky (1965). The restrictions are meant to be hard constraints; as soon as they are violated, the reading is rendered anomalous. Sentence (1.1) gives an example; here, drink requires its direct object to be some kind of liquid.

(1.1) * John drinks the table.

But language is often ambiguous, and in these situations it makes sense to drop the notion of hard constraints in favor of selectional *preferences*. Some examples illustrate the idea:

(1.2) *Peter sitzt auf der Bank.*
Peter sits on the bench/bank.

(1.3) (a) *Peter sieht den Mann mit dem Teleskop.*
Peter sees the man with the telescope.

(b) *Peter sieht den Mann mit dem roten Hut.*
Peter sees the man with the red hat.

(1.4) *Peter mag seinen Hund, obwohl er ihn manchmal beißt.*
Peter likes his dog, although it/he him/it sometimes bites.
'Peter likes his dog, although it bites him/he bites it sometimes.'

The German word *Bank* in sentence (1.2) is lexically ambiguous; it can either mean the bench in a park or the financial institution. Example (1.3) illustrates a prepositional phrase attachment ambiguity; the PPs *mit dem Teleskop* and *mit dem roten Hut* can modify either *Peter* or *den Mann*. Thirdly, the anaphoric pronoun *er*

in sentence (1.4) can refer either to the dog biting Peter, but also to Peter biting the dog.

Selectional restrictions do not help to decide which of the alternatives are preferred; theoretically, all readings are imaginable and would be licensed. For instance, it is possible for men to bite dogs. But as soon as preferences are available, we can choose the park bench, let Peter use the telescope and let the man wear the hat, and decide in favor of the dog biting Peter.

A further use of preferences is to include them in dictionary entries; if learners of a foreign language see the graded selectional profile of a word, it might help them to find the corresponding lexical item in their mother tongue.

1.2. Automatic Acquisition of Selectional Preferences

It is time-consuming and often difficult to assign selectional preference values to predicates manually. Therefore, in the last decade various methods have been suggested to automatically acquire selectional preference information from corpora, large collections of language in use. Additional world knowledge is introduced by utilizing ontologies which provide a structured representation of the meaning of lexical units.

The acquisition methods have been evaluated. However, most evaluations have been conducted for English, only for the verb/direct object relation, and looking at algorithms in isolation. Open questions are whether the methods are valid cross-linguistically, how they perform on other grammatical relations, and how they perform in combination. This information can justify or rebut the usefulness of state of the art techniques for selectional preference acquisition.

The research presented in this thesis attempts to fill these gaps. I conducted a magnitude estimation experiment over the world-wide web to elicit human judgments on the acceptability of 90 German sentences with intransitive and transitive verbs, and verbs subcategorizing for a PP object. These were compared to the preferences computed by five approaches to automatic selectional preference acquisition: Two frequency-based measures, selectional association (Resnik 1993), tree cut models (Li and Abe 1998), and the similarity-class measure (Clark and Weir 2001).

I found that there exist significant linear correlations between the human judgments and different single algorithms, depending on the grammatical relation being observed. Furthermore, when methods are combined by a principal components factor analysis and multiple regression, they outperform the results of a single approach when judging all of the test sentences. This indicates that the approaches are indeed valid for a language other than English, that they work for different grammatical relations, and that it is a promising direction of research to try and combine

approaches to obtain more adequate preference values.

1.3. Structure of the Thesis

In Chapter 2, I describe the methodological foundations of the subsequent work, such as the corpus and parser that were employed, the GermaNet lexical database, and paradigms and statistical methods for evaluation.

Chapter 3 features the approaches to selectional preference acquisition which were evaluated, introducing two frequency-based measures as well as the class-based approaches to compute selectional association (Resnik 1993, 1996), tree cut models (Abe and Li 1996, Li and Abe 1998), and similarity-class probabilities (Clark and Weir 2001, 2002).

Chapter 4 deals with the experiment that I conducted to elicit human judgments on the selectional preferences of a set of experimental stimuli for each of the grammatical relations subject, direct object, and PP object. In Chapter 5, the judgments of the aforementioned algorithms on the verb/argument pairs of the stimuli are evaluated against the human judgments.

Chapter 6 briefly summarizes the work presented in the thesis. Conclusions are drawn and questions for further research in this area are raised.

1. Introduction

2. Methodology

This chapter introduces the methodological foundations of the subsequent work. I describe the corpus (Section 2.1) and parser (Section 2.2) that were employed, the GermaNet lexical database (Section 2.3), and paradigms and statistical methods for evaluation (Section 2.4).

The parser is especially important, as it incorporates a grammatical relation recognizer which extracts the data from which the selectional preferences are acquired. Therefore, I conducted an evaluation of this component, the results of which are presented in Section 2.2.1.

2.1. Corpus

The corpus used for the research reported in this thesis consists of the text of five volumes, 1995 to 1999, of the German newspaper Süddeutsche Zeitung (SZ). It comprises 179 million tokens. The corpus was compiled by the company CLT Sprachtechnologie GmbH and is available to the Saarland University's Department of Computational Linguistics for research purposes.

2.2. The SMES Information Extraction Core System

The algorithms for selectional preference acquisition which are evaluated subsequently require as input co-occurrence triples of the form ⟨*verb, grammatical relation, noun*⟩. To acquire this information I employed SMES, a robust information extraction core system for the processing of German text (Neumann et al. 1997). SMES combines shallow processing techniques, e.g., finite state regular expression recognizers, with generic linguistic resources like a morphology component and a subcategorization dictionary.

SMES incorporates a set of modules to process text. Firstly, a tokenizer maps the text into a stream of tokens. In the stage of lexical processing, the tokens are analyzed morphologically; nominal, adjectival, and verbal compounds are detected,

and part-of-speech tags are assigned. During syntactic processing, the chunk parser module identifies phrases and clauses by means of finite state grammars.

In a third step, verbal grammatical relations (GRs) are recognized. A large subcategorization lexicon is exploited, which contains 11,998 verbs and a total of 30,042 subcategorization frames (Buchholz 1996). It also provides information about verbal arity, case of NP complements, and the various types of sentential complements a verb may take.

2.2.1. Evaluation of the Grammatical Relation Recognizer

To be able to assess the quality of the grammatical relations that are found by SMES, I evaluated the grammatical relation recognizer (GRR). Carroll et al. (1998) propose the annotation of GRs as an alternative to constituency-based parser evaluation. Each sentence in a corpus is marked up with a set of GRs specifying the syntactic dependency between the head and its dependents. Carroll et al. argue that their scheme is application-independent and can deal with language phenomena of English, French, German, and Italian.

The hierarchy of GRs is shown in Figure 2.1. A test corpus of 500 sentences (8,000 words) was randomly sampled from the Süddeutsche Zeitung corpus. Two annotators marked up the sentences, considering only the verbal complex, as the GRR does not take nouns and their complements or modifiers into account.

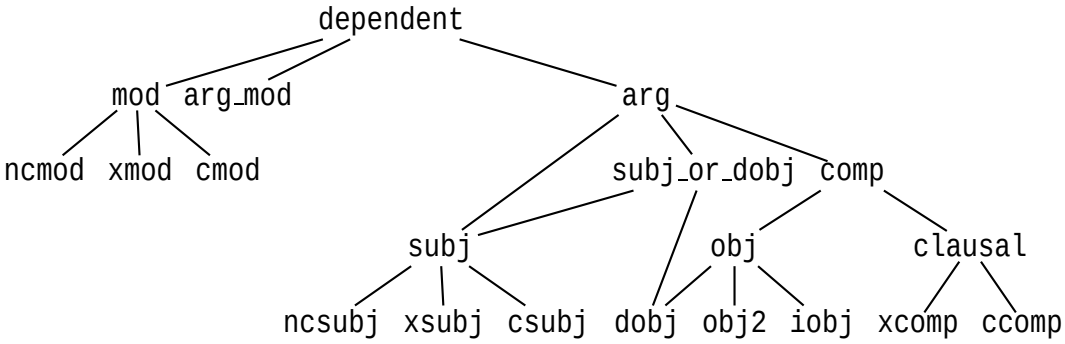

Figure 2.1.: Grammatical relation hierarchy according to Carroll et al. (1998)

The annotators were supplied with Carroll et al. (1998)'s definitions of GRs and a list of annotation guidelines (see Appendix A) which took German-specific phenomena into account (e.g., Zustandspassiv). The annotators were trained on 100 sentences randomly selected from the Berliner Zeitung, another corpus of German newspaper texts (100 million tokens). Markup was done semi-automatically by first

generating the set of relations predicted by the GRR and then manually correcting and extending these. The mean number of GRs per corpus sentence was 3.96 (mean sentence length was 16.94 words).

To give an example, sentence (2.1) was annotated with the GRs given in (2.2). `ncsubj` are non-clausal subjects, `dobj` are direct objects, `cmod` are clausal modifiers, and `mod` is the relation between a head and its modifier; a detailed description of the available GRs is given in Carroll et al. (1998).

(2.1) *Wenn uns mit den Gewerkschaften keine Vereinbarung gelingt,*
if us with the trade unions no agreement achieve,
müssen wir Mitarbeiter entlassen.
must we co-workers lay off
'If we fail to reach an agreement with the trade unions we will have to lay off staff.'

(2.2)
```
ncsubj(entlassen,wir,_)
dobj(entlassen,Mitarbeiter,_)
cmod(wenn,entlassen,gelingen)
ncsubj(gelingen,Vereinbarung)
dobj(gelingen,uns,iobj)
mod(mit,gelingen,Gewerkschaft)
```

Table 2.1 shows the agreement between the two annotators. In Table 2.2, the GRR is compared against the annotated corpus. Precision, recall, and F-score (which combines precision and recall with equal weight) are computed for each type of relation and overall, according to equations (2.3)–(2.5). The figures were obtained using one annotator as the gold standard. The GRR found at least one grammatical relation for 57.7% of the sentences; the figures in Table 2.2 are based on these analyzed sentences.

$$\textit{precision} = \frac{|\textit{found and correct GRs}|}{|\textit{found GRs}|} \tag{2.3}$$

$$\textit{recall} = \frac{|\textit{found and correct GRs}|}{|\textit{correct GRs}|} \tag{2.4}$$

$$\textit{F-score} = \frac{2 \cdot \textit{precision} \cdot \textit{recall}}{\textit{precision} + \textit{recall}} \tag{2.5}$$

The annotators achieved an overall F-score of 86.7%, whereas the GRR reached an overall F-score of 37.8%. Expectedly, precision is better than than recall; the GRR achieved an overall precision of 54.1% and a recall of 29.1% (see row All_1 in Tables 2.1 and 2.2).

Table 2.1.: Inter-annotator agreement on grammatical relations

GR	Correct	Found	F. and C.	Precision %	Recall %	F-score %
ncsubj	766	761	731	96.1	95.4	95.7
dobj	360	362	339	93.6	94.2	93.9
obj2	19	18	16	88.9	84.2	86.5
iobj	108	150	84	56.0	77.8	65.1
xcomp	161	175	139	79.4	86.3	82.7
ccomp	62	52	42	80.8	67.7	73.7
mod	388	321	260	81.0	67.0	73.3
xmod	7	7	7	100.0	100.0	100.0
cmod	52	47	39	83.0	75.0	78.8
arg_mod	12	11	8	72.7	66.7	69.6
All_1	1935	1904	1665	87.4	86.0	86.7
mod_iobj	496	471	409	86.8	82.5	84.6
All_2	1935	1904	1730	90.9	89.4	90.1

Table 2.2.: Accuracy of grammatical relations found by the GRR

GR	Correct	Found	F. and C.	Precision %	Recall %	F-score %
ncsubj	469	294	162	55.1	34.5	42.5
dobj	245	118	59	50.0	24.1	32.5
obj2	12	9	0	0.0	0.0	–
iobj	69	20	15	75.0	21.7	33.7
xcomp	57	34	21	61.8	36.8	46.2
ccomp	34	1	1	100.0	2.9	5.7
mod	260	136	71	52.2	27.3	35.9
xmod	3	2	1	50.0	33.3	40.0
cmod	31	20	12	60.0	38.7	47.1
arg_mod	6	4	3	75.0	50.0	60.0
All_1	1186	638	345	54.1	29.1	37.8
mod_iobj	329	156	91	58.3	27.7	37.5
All_2	1186	638	350	54.9	29.5	38.4

The annotators were having trouble with the iobj GR, the relation between a predicate and a non-clausal complement, for which the lowest agreement was achieved (65.1%, see Table 2.1). Combining mod and iobj as mod_iobj improves results both for the human annotators and the GRR (see row All_2 in Tables 2.1 and 2.2).

The GRR performs poorly at recognizing double objects in dative constructions and the dependencies between predicates and clausal complements with overt subjects (see rows obj2 and ccomp). In particular, multiply nested structures are not handled very accurately; nested clauses sometimes remain undetected or are misrecognized, which has an effect on the overall recall figures.

Another source of errors are mistakes in the preprocessing phase. Separable verb prefixes are not detected reliably by the morphological component. Unknown verb forms are not handled at all. Case ambiguity of phrases influences the selection of subject/object relations. Finally, some errors are due to tagging mistakes.

The subsequent experiment examines selectional preferences for the subject, direct object, and PP object slot of verbs; the relevant GRs are thus ncsubj, dobj, and mod_iobj. Precision, which ranges between 50.0% and 58.3%, is more important than recall, which ranges between 24.1% and 34.5% (see Table 2.2). The figures are high enough to justify usage of the data; of course, they must be taken into consideration when interpreting the statistical methods' output.

2.3. The GermaNet Lexical Database

GermaNet is a lexical-semantic net for the German language (Hamp and Feldweg 1997). It is compatible with the Princeton WordNet (Fellbaum 1998), but it was created from scratch. The lexicon contains verbs, nouns, and adjectives. Word meanings are represented by sets of synonyms or near synonyms, so-called *synsets*. Different senses of a word are denoted by the different synsets it belongs to. Synsets are organized by semantic relations such as hyponymy ('is-a'), meronymy ('is-part-of'), or antonymy.

GermaNet differs from WordNet as far as some design principles are concerned (cf. Hamp and Feldweg 1997). It includes non-lexicalized artificial concept nodes, which allow for a more fine-grained lexical semantic organization. Contrary to WordNet, cross-classification of concepts is an essential feature. Also, there is a special relation between synsets to treat regular polysemy.

The implementations of the selectional preference algorithms evaluated in this thesis make use of the noun taxonomy of GermaNet, and the information encoded in it in terms of the hyponymy/hypernymy relation. The GermaNet noun hierarchy (version 3.0 of January 29, 2001) contains 23,053 noun synsets. It is thus

Table 2.3.: Classification of the correlation coefficient

Correlation Coefficient	Classification
$\|r\| \leq 0.2$	very low correlation
$0.2 < \|r\| \leq 0.5$	low correlation
$0.5 < \|r\| \leq 0.7$	middle correlation
$0.7 < \|r\| \leq 0.9$	high correlation
$0.9 < \|r\| \leq 1$	very high correlation

considerably smaller than that of WordNet (version 1.7), which includes 74,488 synsets; nevertheless, GermaNet is one of the most mature ontologies available for German, and has been used successfully in a variety of other projects.

2.4. Evaluation

2.4.1. Magnitude Estimation

Magnitude estimation is an experimental paradigm commonly used in psychophysics to obtain judgments on sensory stimuli (Stevens 1975). Psycholinguistic studies have shown that this technique is also applicable to the elicitation of linguistic judgments (Gurman Bard et al. 1996, Lapata 2000, Lapata et al. 2001).

Magnitude estimation requires subjects to assign an arbitrary number to a reference sentence, and judge all following stimuli proportionally to the reference value. Thus, subjects are free to choose their own rating scale and are not limited to predefined ordinal scales.

The magnitude estimation paradigm was used in the experiment described in Chapter 4 to capture the judgments of human subjects on a set of test sentences.

2.4.2. Correlation

The purpose of Chapter 5 is to examine the relationship between the judgments of the human subjects and those of the implemented methods. To assess the degree of the correlation between two variables the *correlation coefficient r* is calculated. It takes a value between -1 and 1; its absolute value is usually interpreted according to Table 2.3.

2.4.3. Linear Regression

The statistical method to describe the relationship between two variables mathematically is *regression*. A formula is sought which, given the value of one (independent) variable, can predict the value of the other (dependent) variable.

For the evaluation I am interested in revealing *linear* relationships between variables, i.e., prediction by an equation for a straight line:

$$y = b \cdot x + a \tag{2.6}$$

The regression line is the one with the minimal sum of the squares of the vertical distances[1] of all points to the line. Parameter b is called the *regression coefficient*; its sign indicates whether there is a negative or a positive correlation.

2.4.4. Multiple Linear Regression

Multiple linear regression takes more than one independent variable into consideration when predicting the dependent variable. An equation of the general form shown in (2.7) is calculated:

$$y = b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_n \cdot x_n + a \tag{2.7}$$

In Chapter 5 this method is used to combine approaches for the prediction of the human judgments.

2.4.5. Factor Analysis

Factor analysis is a method to reduce a large number of variables to a small number of hypothetical quantities called *factors*. These factors are created by grouping variables which are highly correlated to each other. The factors themselves have only a low or no correlation to the other factors, i.e., they are supposed to be independent of each other, while explaining as much variance as possible.

The variety of parameters that are explored during the evaluation (see Section 5.1) leads to a large number of higly correlated variables. Factor analysis is applied to reduce these to a set of independent factors before calculating the multiple regression.

[1] Vertical distance denotes the distance parallel to the axis of the dependent variable.

2. Methodology

3. Selectional Preference Acquisition

In this chapter, I introduce the approaches to selectional preference acquisition that are evaluated subsequently. I first describe assumptions common to all methods (Section 3.1). Then I look at two frequency-based measures which rely only on co-occurrence counts observed in the corpus data (Section 3.2).

I proceed with three more sophisticated approaches which assign probabilities to the classes of a noun ontology. The first of them is *selectional association* (Resnik 1993, 1996), which is based on the information-theoretic measure *relative entropy*, capturing the distance between two probability distributions (Section 3.3).

For the second class-based approach, put forward by Abe and Li (1996), Li and Abe (1998), *tree cut models* are computed (Section 3.4). A tree cut model is a horizontal cut through the noun hierarchy, which mirrors the selectional preferences of a verb. The optimal cut is found by means of the *Minimum Description Length* (MDL) principle.

Thirdly, I look at the *similarity-class* measure developed by Clark and Weir (2001, 2002) (Section 3.5). The idea is to find a suitable level of generalization for a noun by traversing the ontology bottom-up, stopping when the probabilities associated with the set of concepts below a node and those of the siblings of that node differ significantly. The resulting class is then used to estimate a probability value for the noun.

3.1. Common Assumptions

As input data, all methods for selectional preference acquisition described below require co-occurrence triples of the form $\langle$*verb, grammatical relation, noun*$\rangle$. These have to be extracted from corpora using a parser with the ability to detect grammatical relations.

For a given verb/noun pair, the algorithms then output a preference value. Depending on the algorithm, various parameters may be set to influence the outcome, e.g., to calculate highest or mean preference values; these parameters will be introduced below.

3.2. Frequency-Based Approaches

The two measures described in this Section rely solely on frequency counts. Unlike the class-based approaches described subsequently, they do not require an ontology.

3.2.1. Frequency

The frequency measure, $freq(v, rel, n)$, is the number of times a noun co-occurs with a verb in a grammatical relation. For instance, if *water* appears 25 times as the object of *drink*, $freq(drink, obj, water) = 25$.

3.2.2. Conditional Probability

The conditional probability $p(v|rel, n)$ of a verb v given a grammatical relation *rel* and a noun n is estimated by relative frequencies as follows:

$$p(v|rel, n) = \frac{p(v, rel, n)}{p(rel, n)} = \frac{freq(v, rel, n)/N}{freq(rel, n)/N} = \frac{freq(v, rel, n)}{freq(rel, n)} \tag{3.1}$$

Here $freq(v, rel, n)$ is the same frequency count as in Section 3.2.1; $freq(rel, n)$ counts how often n appears in relation *rel*, and N is the total number of nouns observed as arguments of *rel*.

Again as an example, if *water* appears 25 times as the object of *drink*, and 50 times as the object of any verb, $p(drink|obj, water) = \frac{freq(drink, obj, water)}{freq(obj, water)} = \frac{25}{50} = 0.5$.

An alternative is to measure the conditional probability $p(n|v, rel)$ of a noun given a verb and a grammatical relation. The estimation differs only as far as the denominator is concerned:

$$p(n|v, rel) = \frac{p(n, v, rel)}{p(v, rel)} = \frac{freq(v, rel, n)/N}{freq(v, rel)/N} = \frac{freq(v, rel, n)}{freq(v, rel)} \tag{3.2}$$

Both types of conditional probability were compared to human judgments, and $p(v|rel, n)$ performed better in general, so I decided to exclude the $p(n|v, rel)$ measure from further investigation.

3.3. Selectional Association

The model of selectional preferences developed by Resnik (1993, 1996) intends to quantify the influence of a predicate on the frequency distribution of its arguments in a probabilistic framework. This introduction is founded upon the descriptions

in Resnik (1996, Section 2) and Manning and Schütze (1999, Section 8.4). The model can capture the relationship between any class of words that imposes semantic constraints on a gramatically dependent phrase, such as verb/subject, verb/direct object, verb/PP object, adjective/noun, or noun/noun (in noun compounds). For the purpose of illustration, I will concentrate on the verb/direct object relation below.

The two central formal notions of the model are selectional preference strength and selectional association. Selectional preference strength is based upon the information theoretic concept *relative entropy* or *Kullback-Leibler (KL) divergence* $D(p\|q)$, which measures the difference between two probability distributions p and q:

$$D(p\|q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{3.3}$$

Intuitively, if p is taken to be the true distribution, and q an approximation of the true distribution, the relative entropy is the amount of information necessary to add to the approximation to make it fit perfectly. The amount of information is usually measured in bits, so log should be read as log to the base 2 here and throughout. $D(p\|q)$ is always greater than or equal to zero, and equal to zero if and only if $p = q$.

The preference model computes probability distributions over the classes of a noun hierarchy of a lexical resource like WordNet or GermaNet. In contrast to dealing with individual nouns, this reduces the number of parameters to be estimated and thus opposes data sparseness problems.

Let $P(C)$ be the overall distribution of direct object noun classes, and $P(C|v)$ the probability distribution of noun classes in the direct object position of verb v. The *selectional preference strength* $S(v)$ of the verb is defined as the relative entropy between these distributions:

$$S(v) = D(P(C|v)\|P(C)) = \sum_{c \in C} P(c|v) \log \frac{P(c|v)}{P(c)} \tag{3.4}$$

$S(v)$ can be understood as the amount of information the predicate, i.e., the verb, carries about its arguments, the direct objects. The greater the difference between the true distribution $P(C|v)$ and the approximation $P(C)$, the greater is the cost of not taking the verb into account.

Selectional preference strength captures the relationship between a verb and the entire noun class hierarchy. The *selectional association* A is defined between a verb v and a *particular* class c:

$$A(v,c) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{S(v)} \tag{3.5}$$

This measure quantifies the relative contribution of class *c* to the overall selectional preference strength. Selectional association values can be positive or negative, expressing preference or dispreference of the respective class.

When looking for a preference value for a verb/noun pair, Resnik suggests to select the noun class which is assigned the highest selectional association value.

3.4. Tree Cut Models

A different method to acquire selectional preference information is proposed by Abe and Li (1996), Li and Abe (1998). Observed co-occurrence triples, which they call case frame instances, are generalized to case frame patterns.

To generalize, conditional probability distributions are estimated for *tree cuts*, partitions of words in a given thesaurus tree. Each leaf node of the thesaurus stands for a noun, and each internal node denotes a noun class, representing all leaf nodes below it. A tree cut is a set of nodes that covers all leaf nodes of the thesaurus tree.

A *tree cut model M* is defined as a pair of a tree cut Γ, which is a set of classes $C_1, C_2, \ldots, C_k$, and a parameter vector θ specifying a probability distribution over the members of Γ. The probabilities sum to 1.

$$M = (\Gamma, \theta) \tag{3.6}$$

$$\Gamma = [C_1, C_2, \ldots, C_k] \tag{3.7}$$

$$\theta = [P(C_1), P(C_2), \ldots, P(C_k)] \tag{3.8}$$

$$\sum_{i=1}^{k} P(C_i) = 1 \tag{3.9}$$

To select the tree cut model that best fits the data, Li and Abe employ the Minimum Description Length (MDL) principle (Rissanen 1978), a principle of data compression and statistical estimation from information theory. A probability model is characterized by the code length in bits required to describe the model itself (*model description length*) and the data observed through it (*data description length*).

A model nearer the root of the thesaurus tree is simpler and fits the data less well than a model nearer the leaves, which is more complex but fits the data better. The best probability model is the one which minimizes the sum of the description lengths.

3.4.1. Computing Description Length

Given a data sample S encoded by a tree cut model $\hat{M} = (\Gamma, \hat{\theta})$ with tree cut Γ and estimated parameters $\hat{\theta}$, the total description length in bits $L(\hat{M}, S)$ is given by equation (3.10):

$$L(\hat{M}, S) = L((\Gamma, \hat{\theta}), S) = L(\Gamma) + L(\hat{\theta}|\Gamma) + L(S|\Gamma, \hat{\theta}) \tag{3.10}$$

The first two addends form the model description length:

$$L(\Gamma) = \log |G| \tag{3.11}$$

$$L(\hat{\theta}|\Gamma) = \frac{k}{2} \times \log |S| \tag{3.12}$$

$L(\Gamma)$ denotes the code length required to identify the cut in the hierarchy; Li and Abe choose to keep this constant, which makes all cuts equally probable. $|G|$ is the cardinality of the set of all possible tree cuts.

The parameter description length $L(\hat{\theta}|\Gamma)$ depends on k, which is the number of classes on the cut Γ, and on the sample size $|S|$.

The final measure is the data description length $L(S|\Gamma, \hat{\theta})$:

$$L(S|\Gamma, \hat{\theta}) = -\sum_{n \in S} \log P_{\hat{M}}(n|v, rel) \tag{3.13}$$

The parameters of the underlying class-based probability model are calculated via maximum likelihood estimation (MLE) by normalizing the frequencies as follows:

$$P_{\hat{M}}(C|v, rel) = \frac{freq(v, rel, C)}{|S|} \tag{3.14}$$

$$\sum_{C \in \Gamma} P_{\hat{M}}(C|v, rel) = 1 \tag{3.15}$$

$freq(v, rel, C)$ is the sum of the frequencies of the nouns in a class observed in relation *rel* to verb v; $|S|$ is the sample size, and Γ is a tree cut. The probability of a given class is distributed equally among the nouns in it:

$$\forall n \in C : P_{\hat{M}}(n|v, rel) = \frac{P_{\hat{M}}(C|v, rel)}{|C|} \tag{3.16}$$

As the number of cuts in a thesaurus tree is exponential in the size of the tree, Li and Abe (1998) devised the algorithm Find-MDL which is based on dynamic programming and efficiently selects the tree cut model with minimum description length. For each child subtree of a given tree, Find-MDL recursively computes the optimal model and appends the results. If the model at the given tree's root has a lower total description length, the lower-level optimal models are collapsed into a model consisting only of the root node.

3.4.2. Adaptations for Use with GermaNet

The GermaNet noun hierarchy is a directed acyclic graph (DAG), but Find-MDL operates on a tree. Following Li and Abe (1998), each subgraph having multiple parents is copied to transform the DAG into a tree. A drawback is that after this modfication, Find-MDL is no longer guaranteed to find the optimal tree cut model.

Furthermore, ambiguous nouns occur at different nodes in the hierarchy. The observed frequency of a noun is thus distributed equally between all nodes of a noun.

A third modification is necessary because in GermaNet, nouns do not only occur as leaves of the hierarchy, but also at internal nodes. When an internal node contains a noun occurring in the observed data, Li and Abe assign the frequencies of all nodes below it to the internal node and delete the subgraph at that position. This way they obtain a *starting cut* as input for the generalization process.

Contrary to this, I decided to follow the practice of Wagner (2000), McCarthy (2001) and created a new leaf for each internal node, containing a copy of the internal node's nouns. This guarantees that all nouns are present at the leaf level.

Lastly, the Find-MDL algorithm requires a single root node for the thesaurus. For WordNet and GermaNet, an artificial concept ⟨`root`⟩ has to be created and connected to the existing top-level classes. WordNet[1] has nine such *unique beginners*, e.g., ⟨`entity`⟩, ⟨`psychological_feature`⟩, or ⟨`abstraction`⟩. From any noun synset below the top-level, the hypernym pointers can be followed to a unique beginner.

On the other hand, GermaNet's[2] noun hierarchy contains 502 synsets without a hypernym. 377 of these have no hyponym, and are thus not linked into the hierarchy by the hyponymy/hypernymy relation at all, but rather by meronymy/holonymy. This leaves 125 "root" classes with no mother node and one or more daughters.

A high number of classes below ⟨`root`⟩ effects a high model description length at this level. Consequently, the generalization process leads to a high amount of tree cuts consisting only of ⟨`root`⟩, which are cheaper because of the lower model description length, but do not offer interesting information about the selectional preferences of a verb. To explore this effect, I set the number of classes below ⟨`root`⟩ as a parameter. Classes which had less than or equal to 10, 20, and 30 hyponyms were excluded from the hierarchy, which left 49, 40, and 33 classes below ⟨`root`⟩, respectively.

[1] Version 1.7

[2] Version 3.0 of 2001-01-29

3.5. Similarity-Class Measure

Unlike the previous two approaches, for which the ontology is crucial to determine a selectional preference profile for a verb, Clark and Weir (2001, 2002) developed a method which is mainly concerned with estimating the probability of a single noun in a given relation to a verb. For this, they also employ a semantic hierarchy, but the main use of it is to overcome the sparse data problem. The idea is to determine an adequate level of generalization in the hierarchy using a chi-square test and to apply this to estimate the probability.

Clark and Weir deal with two questions, which are addressed in the subsequent sections: Firstly, how can a suitably chosen class be used to estimate the probability of a sense? And secondly, how can an adequate class be determined for a noun sense?

3.5.1. Class-Based Probability Estimation

In the following discussion, I adopt Clark and Weir's terminology as far as the semantic hierarchy of WordNet is concerned. A lexicalized concept or sense is referred to as *concept*, a set of senses is called *class*. The synset $\text{syn}(c)$ of a concept c is the set of synonymous words that can be used to denote the concept. For instance, $\text{syn}(\langle\texttt{beverage}\rangle) = \{\textit{beverage, drink, drinkable, potable}\}$. Let $\text{cn}(c) = \{c|n \in \text{syn}(c)\}$ be the set of concepts that can be denoted by noun n.

The hierarchy's edges connect the nodes by the 'direct – isa' relation; the transitive, reflexive closure of that is the 'isa' relation. If c' isa c, then c is a *hypernym* of c' and c' is a *hyponym* of c. The set of concepts dominated by concept c, including c itself, can thus be formalized as $\overline{c} = \{c'|c' \textit{ isa } c\}$. Finally, the probability to be estimated is $p(c|v,rel)$, the probability that some noun n in $syn(c)$ occurs in relation rel to verb v.

Clark and Weir suggest a way to use a set of concepts $\overline{c'}$, where c' is a hypernym of concept c, to estimate $p(c|v,rel)$. They explain that calculating $p(\overline{c'}|v,rel)$ is not a good solution; this probability would be obtained by summing over the concepts in the set, and is likely to be much greater than $p(c|v,rel)$.

$$p(\overline{c'}|v,rel) = \sum_{c'' \in \overline{c'}} p(c''|v,rel) \tag{3.17}$$

Instead, they show that the set of concepts can be used as a condition in the probability $p(v|\overline{c'},rel)$. They prove that this probability can remain constant when moving up in the hierarchy; during the generalization process (Section 3.5.2), the topmost probability which does not differ significantly is sought.

By Bayes' theorem, this probability can be used to compute $p(c|v,rel)$:

$$p(c|v,rel) = p(v|c,rel)\frac{p(c|rel)}{p(v|rel)} \tag{3.18}$$

For example, the probability $p(\langle\texttt{iced\_tea}\rangle|drink,obj)$ could be estimated using $p(drink|\overline{\langle\texttt{iced\_tea}\rangle},obj)$ as follows:

$$p(\langle\texttt{iced\_tea}\rangle|drink,obj) \approx p(drink|\overline{\langle\texttt{iced\_tea}\rangle},obj)\frac{p(\langle\texttt{iced\_tea}\rangle)|obj}{p(drink)|obj} \tag{3.19}$$

To ensure that the estimates form a probability distribution over the concepts of the hierarchy, a normalization factor is introduced. This leads to the final formula for the *similarity-class probability* p_{sc}, where $[c,v,rel]$ denotes the class chosen for concept c in relation rel to verb v, $\hat{p}$ denotes a relative frequency estimate, and C the set of concepts in the hierarchy:

$$p_{sc}(c|v,rel) = \frac{\hat{p}(v|[c,v,rel],rel)\frac{\hat{p}(c|rel)}{\hat{p}(v|rel)}}{\sum_{c'\in C}\hat{p}(v|[c',v,rel],rel)\frac{\hat{p}(c'|rel)}{\hat{p}(v|rel)}} \tag{3.20}$$

The relative frequency estimates are given below; $f(c,v,rel)$ is the number of (n,v,rel) triples in the data in which n is being used to denote c, and V is the set of verbs in the data.

$$\hat{p}(c|rel) = \frac{f(c,rel)}{f(rel)} = \frac{\sum_{v'\in V} f(c,v',rel)}{\sum_{v'\in V}\sum_{c'\in C} f(c',v',rel)} \tag{3.21}$$

$$\hat{p}(v|rel) = \frac{f(v,rel)}{f(rel)} = \frac{\sum_{c'\in C} f(c',v,rel)}{\sum_{v'\in V}\sum_{c'\in C} f(c',v',rel)} \tag{3.22}$$

$$\hat{p}(v|\overline{c'},rel) = \frac{f(\overline{c'},v,rel)}{f(\overline{c'},rel)} = \frac{\sum_{c''\in\overline{c'}} f(c'',v,rel)}{\sum_{v'\in V}\sum_{c''\in\overline{c'}} f(c'',v',rel)} \tag{3.23}$$

To estimate the frequencies of senses, Clark and Weir follow the approach of Li and Abe (1998) (see Section 3.4.2) and distribute the count for each noun in the data equally among all senses of a noun:

$$\hat{f}(c,v,rel) = \sum_{n\in\mathrm{syn}(c)} \frac{f(n,v,rel)}{|\mathrm{cn}(n)|} \tag{3.24}$$

3.5.2. Generalization

Given a concept c in position r of verb v, the generalization procedure determines a suitable *similarity-class* $\overline{c'}$. The hypernym c' of c is referred to as $\text{top}(c, v, rel)$ because it is located at the similarity class's root. The procedure begins at the hierarchy's leaf level by assigning concept c to a variable *top*. Then successive hypernyms of c are assigned to *top* until a node is reached where the probability of the set of concepts dominated by *top* differs significantly from the probabilities of the sets of concepts dominated by *top*'s sister nodes. In that case, *top* is returned as the result of generalization.

A chi-square test is used to determine if $p(v|\overline{c'}, rel)$ changes significantly by moving up a node in the hierarchy. The null hypothesis is that the probabilities $p(v|\overline{c_i}, rel)$ are the same for each child c_i of c'. If there is no significant difference between them, the null hypothesis is accepted and $p(v|\overline{c'}, rel)$ can be taken as an approximation of its child classes. On the other hand, if a significant difference is found, the null hypothesis is rejected and a good approximation cannot be proven.

There are two statistical tests available, the Pearson chi-square statistic χ^2 and the log-likelihood chi-square statistic G^2, where o_{ij} denotes the observed value for the cell in row i and column j, and e_{ij} the expected value, respectively.

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{3.25}$$

$$G^2 = 2 \sum_{i,j} o_{ij} \log_e \frac{o_{ij}}{e_{ij}} \tag{3.26}$$

Clark and Weir discuss which statistic is more adequate for the task at hand; they conclude that there is no common agreement in the literature, and thus this should be decided on a per-application basis. Another parameter to set is the α value which determines the level of significance for the calculated χ^2 or G^2 test statistic. In the subsequent experiment, I follow Clark and Weir's suggestion to compare results across different values of α and choose the one that maximizes performance.

3. Selectional Preference Acquisition

4. Experiment to Elicit Human Judgments on Selectional Preferences

This chapter focuses on the experiment that I conducted to elicit human judgments on the selectional preferences of a set of 90 experimental stimuli. I first present the materials and the design of the experiment (Section 4.1) and continue to describe the experiment's procedure (Section 4.2), participants (Section 4.3), and results (Section 4.4). In the following chapter, the obtained data is compared to computational models of selectional preferences.

4.1. Materials and Design

4.1.1. Co-occurrence Triples

The Süddeutsche Zeitung corpus was parsed using the grammatical relation recognition component of SMES. From the output, co-occurrence triples of the form ⟨*verb, grammatical relation, noun*⟩ were extracted for the three grammatical relations subject, direct object, and PP object.

To reduce the risk of paraphrase ratings being influenced by verb/noun combinations unfamiliar to the experiment's participants, I post-processed the triple data. A triple was only kept if both the verb and the noun were classified in the respective GermaNet hierarchy. Also, verbs and nouns had to appear at least once in a million words of the corpus data.

4.1.2. Construction of Experimental Stimuli

Ten verbs were selected randomly for each grammatical relation. The dependent nouns of each verb were split into three "probability bands" according to frequency.

For each verb, a high, middle, and low frequent dependent noun was chosen randomly.

Therefore, the experimental design consisted of the factors grammatical relation (*Rel*), verb (*Verb*), and probability band (*Band*). The factors *Rel* and *Band* had three levels each, and the factor *Verb* had 10 levels. This yielded a total of $Rel \times Verb \times Band = 3 \times 10 \times 3 = 90$ stimuli. All of the stimuli are listed in Table 4.1 in Section 4.4.

The 90 verb/noun pairs were paraphrased to create sentences. For the direct/PP object sentences, one of 10 common human first names (five female, five male) was added as subject where possible, or else an inanimate subject which appeared frequently according to the corpus data. The stimuli sentences of the verbs *schmieden* and *riechen nach* are shown in (4.1) and (4.2), respectively, sorted by descending frequency of the verb/object pair.

(4.1) (a) *Peter schmiedete einen Plan. [30×]*
Peter forged a plan.

(b) *Peter schmiedete eine Allianz. [8×]*
Peter forged an alliance.

(c) *Peter schmiedete ein Instrument. [1×]*
Peter forged an instrument.

(4.2) (a) *Die Luft roch nach Gas. [4×]*
The air smelled of gas.

(b) *Die Luft roch nach einer Verschwörung. [2×]*
The air smelled of a conspiracy.

(c) *Die Luft roch nach einer Runde. [1×]*
The air smelled of a round.

4.2. Procedure

A magnitude estimation experiment was conducted to obtain judgments on the resulting 90 sentences (cf. Section 2.4.1). The experiment was administered over the Internet. Subjects used their Java enabled web browser to access a server running the WebExp software (Keller et al. 1998). The experiment was self-paced, and response times were recorded to be able to check them for anomalies. A session lasted approximately 20 minutes. The subjects first read a page of instructions and completed a demographic questionnaire. The main experiment consisted of a training phase, a practice phase, and a test phase.

The instructions web page (see Appendix B) contained general information about the experiment and the software prerequisites necessary for participation. Introductory information familiarized the subjects with the concept of magnitude estimation. The upcoming phases of the experiment were described.

In the demographic questionnaire, subjects were asked for their name, e-mail address, age, handedness, job or topic of studies, and the language region in which they grew up. See Section 4.3 for an evaluation of the answers.

During the training phase, subjects were asked to judge the length of five lines relative to a reference line. In the practice phase, they were exposed to a sample reference sentence and six practice stimuli constructed like the ones for the main experiment.

After this preparation, the subjects did the actual experiment. They gave a value to the reference sentence (4.3) and judged the 90 stimuli afterwards. The stimuli were presented in random order, with the constraint that no two verbs with the same subcategorization frame followed each other.

(4.3) *Thomas programmierte das Chaos.*
Thomas programmed the chaos.

4.3. Subjects

Seventy-nine persons participated in the experiment. The subjects were recruited over the Internet by an announcement on the Language Experiments Portal web page, `http://www.language-experiments.org/`, and by postings to relevant newsgroups and mailing lists. Participation was voluntary; six prizes of 25 DM (12.78 €) were drawn after the study had ended. Subjects had to be native speakers of German.

The data of eight subjects were eliminated because they did not complete the experiment. Seven subjects entered arbitrary judgment values. Two participants' data showed response onset times of zero, which indicates a software problem. One subject was not a native speaker of German.

This left 61 subjects for analysis. Of these, 20 subjects were female, 41 male; 51 subjects were right-handed, 10 left-handed. The age of the subjects ranged from 17 to 67 years, the mean was 30.3 years. Twenty-three subjects were linguists or students of linguistics.

4.4. Results

The data were first normalized by dividing each numerical judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. Then the data were transformed by taking the decadic logarithm. This transformation ensures that the judgments are normally distributed and is standard practice for magnitude estimation data (Gurman Bard et al. 1996). All analyses were conducted on the normalized, log-transformed judgments, which are listed together with the stimuli in Table 4.1.

A compromise had to be made for the design of the experiment. For each of the direct object/PP object verbs, a subject had to be chosen, and this does not always fit well (cf. example (4.2)). As a consequence, participants' judgments might have been influenced by the artificially introduced context. This context is not available to the algorithms, which just consider ⟨*verb, grammatical relation, noun*⟩ triple data.

Table 4.1.: Materials for the experiment, with mean human judgments

Verb	Probability Band					
	High		Medium		Low	
beten	Papst	.3303	Pfarrer	.3707	Kreuz	−.7307
durchfallen	Antrag	.1609	Plan	.0033	Reifeprüfung	−.6575
glitzern	Sonne	−.0099	Ferne	−.2623	Tau	.2906
herauskommen	Ergebnis	.0848	Buch	.2273	Zeitung	.2100
knurren	Magen	.2794	Mann	.1240	Knabe	.0750
musizieren	Jugend	.1962	Musiker	.2805	Grundschule	−.0303
protestieren	Mensch	.2064	Plan	−.6858	Ordnung	−.4391
schwappen	Welle	.0622	Bier	.0116	Rock	−.5971
stagnieren	Umsatz	.2816	Preis	.2372	Arbeitslosigkeit	.2088
warten	Welt	.1014	Lohn	−.2110	Veröffentlichung	−.1244
belohnen	Kind	.3453	Kunde	.2364	Anstrengung	.0690
bewirken	Veränderung	.3078	Anstieg	.0858	Beispiel	−.4098
enttäuschen	Erwartung	−.0819	Gast	.3090	Politiker	.2212
erlegen	Tier	.2924	Jahr	−.7191	Gesetz	−.6585
formieren	Widerstand	.1881	Truppe	.2055	Kontur	−.4054
importieren	Rindfleisch	.2474	Japaner	−.2604	Milch	.1108
kürzen	Subvention	.2043	Leistung	.0418	Seite	.1042
pumpen	Geld	−.1541	Tag	−.7024	Leichtigkeit	−.6834
reinigen	Luft	.2178	Tag	−.7146	Gehweg	.2007
schmieden	Plan	.3208	Allianz	.0484	Instrument	−.1471
appellieren an	Seite	−.4434	Zeug	−.4553	Entscheidungsträger	.2676
denken an	Rücktritt	.2895	Freund	.3645	Kleinigkeit	.1897
erkranken an	Brustkrebs	.3580	Malaria	.3549	Durchfall	.3096
grenzen an	Unverschämtheit	.2868	Rassismus	.2822	Betrug	.2733
kommen zu	Schluss	.2199	Urteil	.2743	Wende	−.0710
prallen gegen	Baum	.3570	Berg	.0586	Lastwagen	.3345
riechen nach	Gas	.2482	Verschwörung	.0067	Runde	−.5096
schweigen zu	Vorwurf	.3155	Frage	.2948	Thema	.3290
teilnehmen an	Seminar	.3525	Lehrgang	.3567	Messe	.3368
verzichten auf	Start	.2757	Kandidatur	.3413	Hilfsmittel	.2969

5. Evaluation of the Preference Measures

In the previous chapter, I described the experiment in which human judgments on a set of sentences were elicited. Subsequently, the judgments of the algorithms outlined in Chapter 3 on the verb/argument pairs of these sentences are evaluated against the experiment's data.

Section 5.1 deals with the parameter settings that were varied to create different models for the human judgments. The resulting variables' degree of linear relationship to the human judgments is examined in Section 5.2. To explore the influence of a combination of methods, the variables are reduced to factors which are then used to model the judgments with multiple regression (Section 5.3). The results are discussed (Section 5.4).

5.1. Modeling the Judgments

The methods outlined in Chapter 3 were implemented as Perl scripts.[1] As input, they used the triple data that had already been extracted for the selection of the experiment's materials (cf. Section 4.1.1).

Except for the frequency-based approaches, there was a choice of parameters to set when computing the preference value for a given verb/noun pair, as illustrated in Table 5.1. For selectional association, the choice was between the highest value, as suggested by Resnik, and the mean value over all classes.

In regard to the tree cut models, again highest and mean value were computed, which differed when a noun had more than one parent class on the cut. Secondly, Li and Abe's algorithm requires the ontology to be a directed acyclic graph with a single root node. In contrast to this requirement, as already mentioned in Section 3.4.2, GermaNet's noun hierarchy has 502 synsets without a hypernym class.

[1] For Clark and Weir's algorithm, I adapted an existing implementation by Frank Keller and Mirella Lapata from WordNet to GermaNet. I am grateful for their permission to let me use their source code as a basis.

Table 5.1.: Explored parameter settings

Sel. Assoc.	Tree Cut Model		Similarity-Class			
			highest		mean	
	highest	mean	gscore	chi	gscore	chi
highest, mean	33 c.b.r., 40 c.b.r., 49 c.b.r., 125 c.b.r.		$\alpha = .0005$, $\alpha = .05$, $\alpha = .3$, $\alpha = .75$, $\alpha = .995$			

c.b.r.: classes below root

Amongst these, there are 125 classes with more than one hyponym. A root node was inserted above the topmost 33, 40, 49, and 125 classes which corresponds to excluding classes with less than or equal to 30, 20, 10, and 1 hyponym classes. Thus, the number of classes below root constitutes a further parameter. No pruning techniques were applied to the ontology.

Finally, for Clark and Weir's approach, there was a choice between highest and mean value when a noun was ambiguous, between χ^2 (chi) and G^2 (gscore) statistic for the chi-square test, and between five α values for the respective test's level of significance (.0005, .05, .3, .75, and .995).

All resulting preference values were transformed by taking the decadic logarithm, just like the human judgments (cf. Section 4.4). The only exception are the selectional association values; they are computed using logarithms, so no further transformation is necessary.

5.2. Linear Regression and Correlation Analyses

Regression and correlation analyses were performed to assess the degree of linear relationship between the human judgments as a dependent variable and the algorithms with each of the possible parameter settings, corresponding to 30 different independent variables. I examined the subject, direct object, and PP object sentences in isolation as well as at the 90 sentences altogether. Table 5.2 lists the best correlation coefficients per preference measure, indicating the respective parameters where appropriate. For each grammatical relation, the optimal coefficient is emphasized.

Table 5.2.: Best correlations between human judgments and individual selectional preference measures

Judgment on	Frequency	Cond. Prob.	Sel. Assoc.	Tree Cut Model	Similarity-Class
subject	$r = .386^{*}$	$r = .010$	$\mathbf{r = .408}^{*}$ [highest]	$r = .281$ [mean, 40 c.b.r.]	$r = .268$ [mean, gscore, $\alpha = .75$]
direct object	$r = .360$	$r = .399^{*}$	$r = .430^{*}$ [mean]	$r = .251$ [mean, 40 c.b.r.]	$\mathbf{r = .611}^{***}$ [highest, gscore, $\alpha = .05$]
PP object	$r = .168$	$r = .335$	$r = .330$ [mean]	$r = .319$ [mean, 33 c.b.r.]	$\mathbf{r = .597}^{***}$ [highest, gscore, $\alpha = .3$]
overall	$r = .301^{**}$	$\mathbf{r = .374}^{***}$	$\mathbf{r = .374}^{***}$ [highest]	$r = .341^{***}$ [mean, 40 c.b.r.]	$r = .232^{*}$ [highest, gscore, $\alpha = .3$]

*: $p \leq .05$ **: $p \leq .01$ ***: $p \leq .001$ c.b.r.: classes below root

Table 5.3.: Four retained principal component factors

Factor	Eigenvalue	Difference	Proportion %	Cumulative %
1	7.969	4.718	53.1	53.1
2	3.251	2.065	21.7	74.8
3	1.185	0.333	7.9	82.7
4	0.853	0.215	5.7	88.4

5.3. Factor Analysis and Multiple Linear Regression

A multiple linear regression analysis was employed to explore the combined contribution of methods. Multiple regression is not applicable when too many of the examined variables are highly correlated to each other. Inspection of the data revealed a high degree of multicollinearity. Therefore, independent variables correlating to another one with $r \geq .99$ were dropped, which left over the 15 variables listed in the left column of Table 5.4.

After that, a principal-components factor analysis was performed on all 90 observations, keeping factors that explained more than 5% of the variance, which led to the four factors shown in Table 5.3. Interpretation of the varimax rotated factor loadings (Table 5.4) allowed to assign names to the factors (Table 5.5). These factors were not correlated to each other any more, and could thus be used as predictors in a multiple regression analysis.

Multiple regression on all observations, with all four factors and forward selection with a significance level of $p \geq .05$ for removal from the model, yielded the regression equation (5.1). The corresponding correlation coefficient is $r = .470$ ($p \leq .001$).

$$\text{hum\_judgm} = .091\ \text{f\_cp} + .068\ \text{f\_tcm} + .103\ \text{f\_sa} + .052 \tag{5.1}$$

Factor analyses on the 30 observations for subject, direct object, and PP object, respectively, did not result in factors which were as clearly interpretable as those above; also, predicting the individual relations' judgments with the factors determined from all 90 observations did not lead to an improvement over the predictions with the original un-factorized variables. I assume that this is a consequence of the sparse data available for those computations.

Table 5.4.: Varimax rotated factor loadings

	Factor			
Variable	1	2	3	4
Frequency	.255	.367	.469	.242
Conditional Probability	.115	.104	.165	**.962**
Selectional Association [highest]	.096	.312	**.841**	.112
Selectional Association [mean]	.072	.198	**.920**	.112
Tree Cut Model [highest, 33 c.b.r.]	.110	**.886**	.292	.068
Tree Cut Model [highest, 49 c.b.r.]	.115	**.918**	.220	.053
Tree Cut Model [highest, 125 c.b.r.]	.109	**.852**	.062	.061
Similarity-Class [h., gscore, $\alpha = .0005$]	**.915**	.165	.139	$-.165$
Similarity-Class [h., gscore, $\alpha = .05$]	**.949**	.112	.125	$-.040$
Similarity-Class [h., gscore, $\alpha = .3$]	**.964**	.093	.071	.056
Similarity-Class [h., gscore, $\alpha = .75$]	**.967**	.075	.037	.040
Similarity-Class [h., gscore, $\alpha = .995$]	**.971**	.031	.015	.112
Similarity-Class [h., chi, $\alpha = .0005$]	**.955**	.066	.043	.079
Similarity-Class [h., chi, $\alpha = .05$]	**.970**	.064	.038	.122
Similarity-Class [h., chi, $\alpha = .3$]	**.968**	.056	.040	.135

c.b.r.: classes below root h.: highest

Table 5.5.: Names for the factors

Factor	Name	Abbreviation
1	Similarity-Class	f_sc
2	Tree Cut Model	f_tcm
3	Selectional Association	f_sa
4	Conditional Probability	f_cp

c.b.r.: classes below root

5.4. Discussion

As can be seen from Table 5.2, the preference measures performed differently well for the three GRs in question. Selectional association is the best to model judgments on subjects (low correlation, $r = .408$, $p \leq .05$), closely followed by the simple frequency measure (low correlation, $r = .386$, $p \leq .05$).

The similarity-class method yields middle correlations for the direct object relation (highest value, G^2 statistic, $\alpha = .05$, $r = .611$, $p \leq .001$) as well as for the PP object relation (highest value, G^2 statistic, $\alpha = .3$, $r = .597$, $p \leq .001$).

For all 90 sentences, conditional probability and selectional association (highest value) work equally well (low correlation, $r = .374$, $p \leq .001$), closely followed by the tree cut model (mean value, 40 classes below root, low correlation, $r = .341$, $p \leq .001$).

Interestingly, there is no single method which outperforms all the others; each algorithm has its strengths and weaknesses. Also, the more sophisticated class-based approaches to selectional preference acquisition do not always achieve better results than the frequency-based ones which do not use an ontology.

It is noticeable that all measures are positively correlated with the human judgments, although several of them do not reach significance. Especially, conditional probability cannot predict subject preferences ($r = .010$), and the frequency measure is not suitable for PP objects ($r = .168$).

As far as parameter settings are concerned, tree cut models seem to work best with the mean preference value and 40 classes below root, and the similarity-class measure yields optimal results using the highest value and the gscore statistic. On the other hand, an interesting result of the factor analysis is that parameter settings do not seem to play an important role—every factor corresponds to a different measure.

Multiple regression with the four factors on all 90 observations results in an improved model to predict the human judgments; although still low, the correlation coefficient rises from .374 ($p \leq .001$) to .470 ($p \leq .001$). In the regression equation (5.1), the factors for conditional probability, tree cut model, and selectional association have been combined to achieve this outcome, which indicates that the combination of approaches can enhance results.

This observation has to be interpreted with care, as the computed model was applied to the data it was trained on. Ideally, the model would be checked against unseen human judgments, but that was not possible in this case. Nevertheless, the function resulting from multiple linear regression is a very simple one, and even with this simple function an improvement on the training data was achieved. Thus, a similar improved performance can be expected for unseen data.

The preference measures depend on the quality of the $\langle$*verb, grammatical rela-*

tion, noun⟩ co-occurrence triple data from which they are calculated. The evaluation in Section 2.2.1 indicated a precision of 55.1% for the subject triples, 50.0% for the direct object data, and 58.3% for the PP object triples (cf. Table 2.2). This is adequate to achieve the degree of correlation reported above. In spite of that, more precise input triples are likely to lead to higher correlations.

5. Evaluation of the Preference Measures

6. Conclusions

6.1. Summary of Contributions

In this thesis, approaches to the automatic acquisition of selectional preference information were evaluated. Previous evaluations have been conducted for the English language, which mainly concentrated on examining the preference relation between a verb and its direct object only, and considered algorithms in isolation. I evaluated the algorithms for the German language, looked into the three grammatical relations subject, direct object, and PP object, and explored a method to combine approaches.

A magnitude estimation experiment was administered over the Internet which resulted in human judgments on 30 sentences for each of the three grammatical relations. Two frequency-based and three class-based preference measures (selectional association (Resnik 1993), tree cut models (Li and Abe 1998), and the similarity-class measure (Clark and Weir 2001)) were implemented and used to compute preference values for the experiment's items, exploring a variety of parameter settings.

Linear regression and correlation analyses showed that there exist significant correlations between the human judgments and the individual selectional preference measures, which demonstrates that the approaches are applicable cross-linguistically. The findings showed that the measures were suited differently well to predict the respective grammatical relations.

A factor analysis was performed to reduce the number of models that resulted from the large number of parameters. Each of the four resulting factors clearly corresponded to one of the algorithms. The factors were combined using multiple linear regression, and the resulting model correlated better to the judgments on all 90 sentences than any individual algorithm.

6.2. Further Research

It would be interesting to repeat the web-based experiment with a different set of stimuli and examine whether the correlation results are reproducible. Then there

would also be the possibility to train a multiple regression model on one set of human judgments and check the performance when applying it to the other data set.

The correlations found in the evaluation of the approaches to selectional preference acquisition were reasonably high, but there is still room for improvement. There are various directions of research that can be pursued to enhance the computed preference values.

The findings of the thesis indicate that it is promising to combine algorithms, as each of them has different strengths and weaknesses. It would be interesting to explore other machine-learning techniques for numeric prediction and see how they compare to multiple linear regression.

It can be expected that the preference values become more adequate if the quality of the input data is improved, so other ways for grammatical relation recognition could be investigated. An alternative for German might be the similarity-based algorithm proposed by Kübler and Hinrichs (2001), which assigns functional labels to complete syntactic structures on the basis of pre-chunked input; in an evaluation, functional labels were recognized with a correctness of 89.73%.

Moreover, the acquisition approaches themselves can be modified. The evaluated methods rely only on the $\langle$*verb, grammatical relation, noun*$\rangle$ triple data when computing the preference values. It would be interesting to develop and explore a model which takes discourse context into account.

A. Annotation Guidelines for the SMES Evaluation

The SMES grammatical relation recognizer was evaluated using a subset of the grammatical relations (GRs) suggested by Carroll et al. (1998) (see Section 2.2.1). A set of test sentences was annotated manually. The annotation guidelines follow.

Grammatical Relations

We are interested in ten GRs. If they exist, they should be listed in the order given below, first for the matrix sentence, then for embedded sentences and relative clauses. Use one line per GR.

- ncsubj(*head*,*dependent*,*initial_gr*)
- dobj(*head*,*dependent*,*initial_gr*)
- iobj(*type*,*head*,*dependent*)
- obj2(*head*,*dependent*)
- xcomp(*type*,*head*,*dependent*)
- ccomp(*type*,*head*,*dependent*)
- mod(*type*,*head*,*dependent*)
- xmod(*type*,*head*,*dependent*)
- cmod(*type*,*head*,*dependent*)
- arg_mod(*type*,*head*,*dependent*,*initial_gr*)

Annotation Guidelines

See Carroll et al. (1998) for English example GR instantiations. Note the following aspects:

- verbal heads are to be specified in their infinitive form, including a possibly separated prefix or *sich* for reflexive verbs
- in constructions with modal verbs, annotate the infinitive as verbal head
- nouns are to be specified in singular; lower/upper case does not matter
- in relative sentences, annotate the relative pronouns as dependents; do not resolve the anaphor (e.g., *Der Mann, der singt, ...* ↦ `ncsubj(singen,der,_)`)
- in cases of apposition, choose the proper name (e.g., *CDU-Generalsekretärin Angela Merkel* ↦ *Angela Merkel*)
- for proper names, annotate first and last name; for location names, choose two words when the last one does not "convey the meaning" (e.g., *Regionalflughafen Augsburg*, *Straße 90*)
- expand words abbreviated by hyphenation (e.g., *Kommissions- und Kostensätze* ↦ *Kommissionssatz*, *Kostensatz*)
- for passive sentences, we are looking for the subject before any GR-changing process, i.e., something like `ncsubj(`*`head`*`,`*`dependent`*`,obj)`
- annotate *Zustandspassiv* like other passive constructions (e.g., *Die Quote ist zu bezahlen.* ↦ `ncsubj(bezahlen,Quote,obj)`)
- non-overt subjects of clausal complements or elliptical constructions are to be annotated
- accusative as well as dative objects are analyzed as `dobj`; the latter are marked with the `initial_gr` of `iobj`
- `obj2` is appropriate only for the second argument of ditransitive verbs
- annotate non-clausal modifiers with `mod` (not `ncmod`); use `xmod` or `cmod` for the clausal ones
- only annotate verbal modifiers with prepositions (i.e., no NP modifiers and nothing with an empty first slot of `mod`)

- when a PP is a verb's complement, analyze it as `iobj`; else take `mod`
- prepositions are to be specified just like they are used; do not reduce them to a base form (i.e., do not change *im* to *in* or *zum* to *zu*)
- the `type` of constructions with the infinitive conjunction *um zu* is `um zu` (and not only `um` or `zu`); other constructions are treated analogously (e.g., *bis zu*, *innerhalb von*, *über hinaus*)
- for *sein* or *werden* annotate `xcomp`, not `dobj` (see Carroll et al. 1998, p. 7)

A. Annotation Guidelines for the SMES *Evaluation*

B. Instructions for the Web-Based Experiment

The instructions given below were displayed to the participants of the experiment described in Chapter 4.

Experiment: Satzbeurteilung

Geld zu gewinnen!

Vielen Dank für Ihr Interesse an unserem Versuch! Durch die Teilnahme an diesem Experiment nehmen Sie automatisch an unserer **VERLOSUNG** teil. Wenn diese Studie abgeschlossen ist, werden wir unter allen Teilnehmern **sechs Gewinner** auslosen, die **jeweils** einen Scheck von **25 DM** (12,78 €) erhalten. Bitte stellen Sie sicher, dass Sie Ihre E-Mail-Adresse in dem dafür vorgesehenen Feld korrekt angeben. Mehrmalige Teilnahme an dem Experiment ist ausgeschlossen.

Bitte lesen Sie die untenstehende Anleitung sorgfältig durch, bevor Sie mit dem Experiment beginnen. Wenden Sie sich an den Versuchsleiter, falls Sie Fragen haben sollten.

Der Versuch erfordert einen Java-kompatiblen Browser. Java muss eingeschaltet sein. Falls technische Probleme auftreten sollten, ziehen Sie bitte unsere Seite mit technischen Hinweisen zu Rate (derzeit nur auf Englisch verfügbar).

Wenn Sie Netscape einsetzen, verändern Sie bitte nicht die Größe des Browser-Fensters, während das Experiment läuft; ansonsten wird das Java-Programm abgebrochen und kann nicht mehr weiterlaufen. Vergrößern oder verkleinern Sie in diesem Fall zuerst das Fenster und drücken dann [Shift] zusammen mit dem [Neu laden]- bzw. [Reload]-Knopf in der Navigations-Symbolleiste, um das Experiment neu zu starten.

Persönliche Daten

Zu Beginn des Versuchs möchten wir Sie bitten, einige persönliche Daten in ein dafür vorgesehenes Fenster einzutragen. Dieses Fenster erscheint, nachdem Sie das Start-Feld (siehe unten) betätigt haben.

Wir bitten Sie, die folgenden Angaben zu machen:

- Name und E-Mail-Adresse
- Alter und Geschlecht
- ob Sie Rechts- oder Linkshänder sind
- Ihr derzeitiger Beruf (bzw. das Fach, das Sie studieren oder studiert haben)
- „Sprachregion“, d. h. die Gegend, in der Sie *aufgewachsen* sind (Bundesland/Kanton und Stadt)

Das Experiment dient rein wissenschaftlichen Zwecken. Ihre persönlichen Daten werden *streng vertraulich* behandelt und in keinem Fall an Dritte weitergegeben werden. Ihre Antworten werden bei der weiteren Auswertung nicht mit Ihrem Namen in Verbindung gebracht.

Versuchsanleitung

Abschnitt 1: Beurteilung von Linien

Bevor Sie mit dem eigentlichen Versuch beginnen, werden Sie eine kurze Trainingsphase durchlaufen, in der es darum geht, die Länge von Linien zu beurteilen. Sie werden nacheinander eine Reihe von Linien auf dem Bildschirm sehen, deren Länge Sie abschätzen sollen, indem Sie jeder Linie eine Zahl zuweisen. Dabei kommt es darauf an, dass Sie die Länge jeder Linie im Verhältnis zur *Vergleichslinie* beurteilen. Als Vergleichsslinie dient die erste Linie, die angezeigt wird. Dieser Linie können Sie eine beliebige Zahl zuweisen, die dann als Vergleichswert dient.

Nachdem Sie den Vergleichswert bestimmt haben, sollen Sie jeder weiteren Linie eine Zahl zuweisen, die ausdrückt, wie lang diese Linie *im Verhältnis* zur Vergleichslinie ist. Je länger die Linie ist, desto größer wird Ihre Zahl sein. Wenn sie beispielsweise meinen, dass eine Linie zweimal so lang ist wie die Vergleichslinie, dann nehmen Sie den Vergleichswert mal zwei; wenn sie nur ein Drittel so lang ist, dann teilen Sie den Vergleichswert durch drei.

Nehmen wir an, die folgende Linie ist Ihre Vergleichslinie, und Sie weisen ihr beispielsweise die Zahl 10 zu:

Als nächstes sollen Sie die folgende Linie beurteilen. Vielleicht geben Sie ihr eine 17, weil sie fast zweimal so lang ist wie die Vergleichslinie:

Und für die folgende Linie scheint Ihnen 2.5 angemessen:

Für Ihre Beurteilung können Sie beliebige Zahlen verwenden. Sowohl ganze Zahlen als auch Kommazahlen sind zulässig. Wenn Sie z. B. der Vergleichslinie einen Wert von 1 zugewiesen haben, dann sollten sie der letzten Linie eine 0.25 geben. Wichtig ist, dass jede Zahl der Länge der zugehörigen Linie entspricht.

Abschnitte 2 und 3: Satzbeurteilung

Im ersten Abschnitt ging es darum, die Länge von Linien mit Hilfe von Zahlen abzuschätzen. In den Abschnitten 2 und 3 sollen Sie nun die Akzeptabilität von Sätzen auf die gleiche Art und Weise abschätzen.

Es wird Ihnen eine Reihe von Sätzen dargeboten werden, die recht unterschiedlich sind: Einige Sätze sind völlig problemlos, andere klingen mehr oder weniger eigenartig oder sind vielleicht sogar von sehr geringer Akzeptabilität. Ihre Aufgabe besteht darin, zu beurteilen, wie *akzeptabel* diese Sätze sind, indem Sie jedem Satz eine Zahl zuweisen.

Wie im ersten Abschnitt wird zuerst ein Satz als *Vergleichssatz* angezeigt, dem Sie dann einen Vergleichswert zuweisen. Als Vergleichswert können Sie eine beliebige Zahl verwenden. Jedem weiteren Satz weisen Sie dann eine Zahl zu, die angibt, wie akzeptabel oder unakzeptabel dieser Satz *im Verhältnis* zum Vergleichssatz ist.

Nehmen wir an, der Vergleichssatz ist:

(B.1) Martin fütterte die Uhr.

Diesem Satz werden Sie vermutlich eine recht niedrige Zahl zuweisen. Der nächste Satz könnte dann lauten:

(B.2) Die Erde bebte.

Wenn dieser Satz Ihnen im Verhältnis zum Vergleichssatz zehnmal so akzeptabel vorkommt, dann werden Sie ihm eine Zahl zuweisen, die zehnmal so groß ist wie die Vergleichszahl. Wenn der Satz Ihnen halb so akzeptabel erscheint wie der Vergleichssatz, dann teilen Sie die Vergleichszahl durch zwei.

Ihre Zahlen können beliebig klein oder groß sein; auch Kommazahlen sind zulässig. (Aber bitte vermeiden Sie Null oder negative Zahlen.) Versuchen Sie, einen breiten Zahlenbereich zu verwenden und möglichst feine Abstufungen vorzunehmen.

Es gibt hierbei keine „richtigen“ oder „falschen“ Antworten! Beachten Sie zudem, dass wir hauptsächlich an Ihrem *ersten Eindruck* von den Sätzen interessiert sind. Arbeiten Sie die Sätze also bitte *zügig* der Reihe nach durch.

Versuchsablauf

Als Erstes tragen Sie bitte die persönlichen Daten in das dafür vorgesehene Fenster ein, das nach Betätigen des Start-Feldes (siehe unten) erscheint.

Nachdem Sie die persönlichen Daten angegeben haben, bearbeiten Sie bitte nacheinander die folgenden drei Abschnitte:

- Trainingsphase: Beurteilung von 5 Linien
- Übungsphase: Beurteilung von 6 Sätzen
- Experimentalphase: Beurteilung von 90 Sätzen

In jedem Abschnitt wird zuerst die Vergleichslinie oder der Vergleichssatz angezeigt. Bitten geben Sie dann Ihre Vergleichszahl ein und betätigen Sie das Weiter-Feld. Dann werden nacheinander alle weiteren Linien bzw. Sätze angezeigt. Bitte geben Sie Ihre Beurteilung jeweils in das Feld unter der Linie bzw. dem Satz ein. Nun müssen Sie die Return-Taste betätigen, um den nächsten Satz zu erhalten.

Der Versuch wird ca. 20-25 Minuten dauern. Danach werden die Daten automatisch zum Versuchsleiter übertragen und Sie erhalten eine Bestätigung per E-Mail.

Bitte beachten Sie:

- Weisen Sie dem Vergleichssatz eine beliebige Zahl zu.
- Beurteilen Sie die Akzeptabilität jedes weiteren Satzes im Verhältnis zum Vergleichssatz.
- Verwenden Sie beliebige positive Zahlen für Ihre Beurteilung.
- Verwenden Sie große Zahlen für akzeptable Sätze und kleine Zahlen für unakzeptable Sätze. Zahlen im Zwischenbereich sind für Sätze von mittlerer Akzeptabilität zu verwenden.
- Versuchen Sie, einen breiten Zahlenbereich zu verwenden und möglichst feine Abstufungen vorzunehmen.

- Bearbeiten Sie den Versuch zügig und beurteilen Sie die Sätze nach Ihrem ersten Eindruck.

B. Instructions for the Web-Based Experiment

Bibliography

Abe, N. and H. Li (1996). Learning word association norms using tree cut pair models. In *Proceedings of the 13th International Conference on Machine Learning (ICML-96)*, Bari, Italy, pp. 3–11.

Buchholz, S. (1996). Entwicklung einer lexikographischen Datenbank für die Verben des Deutschen. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany.

Carroll, J., T. Briscoe, and A. Sanfilippo (1998). Parser evaluation: A survey and a new proposal. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-98)*, Granada, Spain, pp. 447–454.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA, USA: MIT Press.

Clark, S. and D. Weir (2001). Class-based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, Pittsburgh, PA, USA, pp. 95–102.

Clark, S. and D. Weir (2002). Class-based probability estimation using a semantic hierarchy. *Computational Linguistics 28*(2), 187–206.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press.

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

Hamp, B. and H. Feldweg (1997). GermaNet – a lexical-semantic net for German. In *Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications at the 35th Annual*

Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL-97), Madrid, Spain, pp. 9–15.

Katz, J. J. and J. A. Fodor (1963). The structure of a semantic theory. *Language 39*(2), 170–210.

Keller, F., M. Corley, S. Corley, L. Konieczny, and A. Todirascu (1998). WebExp: A Java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh, UK.

Kübler, S. and E. W. Hinrichs (2001). From chunks to function-argument structure: A similarity-based approach. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL-01)*, Toulouse, France, pp. 338–345.

Lapata, M. (2000). *The Acquisition and Modeling of Lexical Knowledge: A Corpus-based Investigation of Systematic Polysemy*. Ph. D. thesis, University of Edinburgh, UK.

Lapata, M., F. Keller, and S. McDonald (2001). Evaluating smoothing algorithms against plausibility judgements. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL-01)*, Toulouse, France, pp. 346–353.

Li, H. and N. Abe (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics 24*(2), 217–244.

Manning, C. D. and H. Schütze (1999). Lexical acquisition. In *Foundations of Statistical Natural Language Processing*, Chapter 8, pp. 265–314. Cambridge, MA, USA: MIT Press.

McCarthy, D. (2001). *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph. D. thesis, University of Sussex, UK.

Neumann, G., R. Backofen, J. Baur, M. Becker, and C. Braun (1997). An information extraction core system for real world German text processing. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, Washington, DC, USA.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition 61*, 127–159.

Resnik, P. S. (1993). *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph. D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica 14*, 465–471.

Stevens, S. S. (1975). *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. New York, NY, USA: Wiley.

Wagner, A. (2000). Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *Proceedings of the 1st Workshop on Ontology Learning (OL-00) at the 14th European Conference on Artificial Intelligence (ECAI-00)*, Berlin, Germany, pp. 37–42.