Evaluating Approaches to Selectional Preference Acquisition against German Plausibility Judgments

Carsten Brockmann
Institute for Communicating and Collaborative Systems (ICCS)
School of Informatics, The University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
Carsten.Brockmann@ed.ac.uk

Abstract

Previous research on the automatic acquisition of selectional preference information has mainly focused on English and on the relation between verbs and their direct objects. In this paper, we evaluate the performance of models of selectional preferences for German verbs and take into account not only direct objects, but also subjects and prepositional complements. A variety of parameter settings are explored. The preference values are compared to human judgments elicited in a magnitude estimation experiment. The results indicate that there exist significant linear correlations between the human judgments and different single algorithms, depending on the grammatical relation being observed.

1 Introduction

Selectional preferences are graded constraints that a predicate imposes on its arguments. For example, a verb like *drink* typically takes an animate entity as its subject and a drinkable entity as its object. Selectional preferences have been studied in the context of a variety of natural language processing tasks, such as word-sense disambiguation (McCarthy et al., 2001), PP attachment ambiguity resolution (Li and Abe, 1998), parse ranking (Bikel, 2000), and the interpretation of compound nouns (Lapata, 2000).

Results on the usefulness of preference knowledge so far have been inconclusive: Resnik (1997) and McCarthy et al. (2001) argue that selectional preferences can not be the only means for word-sense disambiguation; they only help when the tie between predicate and argument is strong enough, so coverage is limited. Bikel (2000) describes a statistical model for simultaneous syntactic parsing and generalized word-sense disambiguation. He finds that integrating word sense information from WordNet does not improve the parsing accuracy significantly over a traditional parsing model.

In contrast to these task-based evaluations, the present paper compares the computed selectional preferences to human judgments obtained in a magnitude estimation experiment over the world-wide web. This approach is more direct than word-sense disambiguation, which relies on the assumption that models of selectional preferences have to infer the appropriate semantic class and therefore perform disambiguation as a side effect.

Furthermore, previous research on automatic selectional preference acquisition has mainly focused on English and on the relation between verbs and their direct objects. In this study, we explore the cross-linguistic applicability of the models by looking at a set of German verbs, and take into account not only direct objects, but also subjects and prepositional complements.

We implemented the following methods to acquire selectional preference values: Co-occurrence frequency, conditional probability, and three distinct approaches that assign probabilities to the classes of a noun ontology (in our case GermaNet (Hamp and Feldweg, 1997)).

The selectional association metric (Resnik, 1993) is based on the information-theoretic measure of *relative entropy*, capturing the distance between two probability distributions.

For the second class-based approach, *tree cut models* are computed (Li and Abe, 1998). 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 principle.

Third, we considered the *similarity-class* measure (Clark and Weir, 2002). 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.

The evaluation showed that there exist significant linear correlations between the human judgments and different single algorithms, depending on the grammatical relation being observed. This indicates that the approaches are indeed valid for a language other than English and that they work for grammatical relations other than direct object.

The remainder of this paper is organized as follows: In Section 2, we briefly review the methods employed to compute selectional preference information. Section 3 deals with the experiment that we conducted to elicit human judgments for a set of experimental stimuli for each of the grammatical relations subject, direct object, and PP object. In Section 4, we describe necessary adaptations to the taxonomy and the parameter settings we explored to model the judgments. Section 5 presents the results of the comparison between the human judgments and the algorithms' predictions. We discuss the results in Section 6 and indicate topics for further research in Section 7.

2 Methods for Selectional Preference Acquisition

2.1 Co-occurrence Frequency

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

2.2 Conditional Probability

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

$$(1)\qquad \begin{aligned}\hat{P}(n|v,r) &= \frac{\hat{P}(v,r,n)}{\hat{P}(v,r)} = \frac{f(v,r,n)/N}{f(v,r)/N}\\ &= \frac{f(v,r,n)}{f(v,r)}\end{aligned}$$

Here $f(v,r,n)$ is the same frequency count as in Section 2.1; $f(v,r)$ counts how often v and r co-occur, and N is the total number of nouns observed as arguments of r.

For example, if *water* occurs 25 times as the object of *drink*, and *drink* has 50 objects attested in the corpus, $\hat{P}(water|drink, obj) = \frac{f(drink,obj,water)}{f(drink,obj)} = \frac{25}{50} = 0.5$.

Equation (1) can be construed as a verb selecting for a noun. An alternative is to have the argument select for its predicate by measuring the conditional probability $\hat{P}(v|r,n)$ of a verb given a grammatical relation and a noun:

$$(2)\qquad \hat{P}(v|r,n) = \frac{f(v,r,n)}{f(r,n)}$$

2.3 Selectional Association

Resnik (1993) was the first to propose a model of selectional preferences to quantify the semantic fit of a particular semantic class c as an argument of a verb v.

The preference model computes probability distributions over the classes of a hierarchy of a lexical resource like WordNet or GermaNet. Let $P(c)$ be the overall distribution of classes, and $P(c|v,r)$ the probability distribution of argument classes in relation r to a particular verb v. The *selectional preference strength* $S(v)$ of the verb is defined as the relative entropy between these distributions:

$$(3)\qquad S(v) = \sum_{c \in C} P(c|v,r) \log \frac{P(c|v,r)}{P(c)}$$

$S(v)$ can be understood as the amount of information the predicate carries about its arguments. The greater the difference between the true distribution $P(c|v,r)$ 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 argument class hierarchy. The *selectional association A* is defined between a verb v and a *particular* class c:

$$(4)\qquad A(v,r,c) = \frac{P(c|v,r) \log \frac{P(c|v,r)}{P(c)}}{S(v)}$$

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.

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

$$(5)\qquad \hat{P}(c|v,r) = \frac{f(v,r,c)}{f(v,r)}$$

The estimation of $P(c|v,r)$ would be a straightforward task if each word was always represented in the taxonomy by a single concept or if we had a corpus labeled explicitly with taxonomic information. Lacking such a corpus we need to take into consideration the fact that words in a taxonomy may belong to more than one conceptual class. Counts of verb-argument configurations are constructed for each conceptual class by dividing the contribution of the argument by the number of classes it belongs to (Resnik, 1993):

$$\hat{f}(v,r,c) = \sum_{n \in \mathrm{syn}(c)} \frac{f(v,r,n)}{|\mathrm{cn}(n)|} \tag{6}$$

Here, $\mathrm{syn}(c)$ is the synset of a concept c, i.e., the set of synonymous words which can be used to denote the concept (e.g., $\mathrm{syn}(\langle\texttt{beverage}\rangle) = \{$*beverage, drink, drinkable, potable*$\}$), and $\mathrm{cn}(n)$ is the set of concepts that can be denoted by noun n (more formally, $\mathrm{cn}(n) = \{c | n \in \mathrm{syn}(c)\}$).

2.4 Tree Cut Models

A different method to acquire selectional preference information is proposed by Li and Abe (1998). Conditional probability distributions are estimated for *tree cuts*, partitions of words in a given hierarchy tree. Each leaf node of the hierarchy 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 hierarchy 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 one.

$$M = (\Gamma, \theta) \tag{7}$$

$$\Gamma = [c_1, c_2, \ldots, c_k] \tag{8}$$

$$\theta = [P(c_1), P(c_2), \ldots, P(c_k)] \tag{9}$$

$$\sum_{i=1}^{k} P(c_i) = 1 \tag{10}$$

To select the tree cut model that best fits the data, Li and Abe employ the Minimum Description Length 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 hierarchy 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.

Given a data sample S, encoded by the 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 computed by equation (11):

$$\begin{aligned} L(\hat{M},S) = \log|G| + \frac{k}{2}\log|S| \\ - \sum_{n \in S} \log P_{\hat{M}}(n|v,r) \end{aligned} \tag{11}$$

$|G|$ denotes the cardinality of the set of all possible tree cuts, k is the number of classes on the cut Γ, and $|S|$ is the sample size.

The probability of a noun, $P_{\hat{M}}(n|v,r)$, is estimated by distributing the probability of a given class equally among the nouns in it:

$$\forall n \in c : P_{\hat{M}}(n|v,r) = \frac{P_{\hat{M}}(c|v,r)}{|c|} \tag{12}$$

2.5 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 (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.

2.5.1 Class-Based Probability Estimation

Let c' denote a hypernym of concept c, and $\overline{c'}$ the set of concepts dominated by concept c', including c' itself. Clark and Weir suggest a way to use $\overline{c'}$ to estimate $P(c|v,r)$. They explain that calculating $P(\overline{c'}|v,r)$ 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 $P(c|v,r)$:

$$P(\overline{c'}|v,r) = \sum_{c'' \in \overline{c'}} P(c''|v,r) \tag{13}$$

Instead, they show that the set of concepts can be used as a condition in the probability $P(v|\overline{c'},r)$. They prove that this probability can remain constant when moving up in the hierarchy; during the generalization process (see Section 2.5.2), the topmost probability which does not differ significantly is sought.

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

$$P(c|v,r) = P(v|c,r)\frac{P(c|r)}{P(v|r)} \tag{14}$$

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}:

$$P_{sc}(c|v,r) = \frac{\hat{P}(v|[c,v,r],r)\frac{\hat{P}(c|r)}{\hat{P}(v|r)}}{\sum_{c' \in C}\hat{P}(v|[c',v,r],r)\frac{\hat{P}(c'|r)}{\hat{P}(v|r)}} \tag{15}$$

$[c,v,r]$ denotes the class chosen for concept c in relation r to verb v, $\hat{P}$ denotes a relative frequency estimate, and C the set of concepts in the hierarchy. Again, since we are not dealing with word sense disambiguated data, counts for each noun are distributed evenly among all senses of the noun (see equations (5) and (6)).

2.5.2 Generalization

Given a concept c in position r of verb v, the generalization procedure determines a suitable *similarity-class* $\overline{c'}$. 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 $P(v|\overline{c'},r)$ changes significantly by moving up a node in the hierarchy. The null hypothesis is that the probabilities $P(v|\overline{c_i},r)$ are the same for each child c_i of c'. If there is no significant difference between them, the null hypothesis is accepted and $P(v|\overline{c'},r)$ 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. 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, we follow Clark and Weir's suggestion to compare results across different values of α and choose the one that maximizes performance.

3 Eliciting Human Judgments on Selectional Preferences

3.1 Materials and Design

3.1.1 Co-occurrence Triples

The research reported in this paper was conducted on a corpus consisting of the text of five volumes (1995 to 1999) of the German newspaper Süddeutsche Zeitung. It comprises 179 million tokens.

The corpus was parsed using SMES, a robust information extraction core system for the processing of German text (Neumann et al., 1997). This system 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 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.

From the output of SMES, co-occurrence triples of the form $\langle v,\ r,\ n\rangle$ were extracted for the three grammatical relations subject, direct object, and PP object. In order to reduce the risk of ratings being influenced by verb/noun combinations unfamiliar to the participants, we removed triples that had a verb or a noun with frequency less than one per million.

We conducted an evaluation of the grammatical relation recognition component of SMES which 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.

3.1.2 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.

The 90 verb/noun pairs were paraphrased to create sentences. For the direct/PP object sentences, one of ten 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 verb *schmieden* are shown in (16), sorted by descending corpus frequency of the verb/object pair, which is given in brackets after the sentence.

(16) (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.

3.2 Procedure

A magnitude estimation experiment was conducted to obtain judgments on the resulting 90 sentences. 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 pre-defined ordinal scales.

The experiment was administered over the Internet. The participants used their Java enabled web browser to access a server running the WebExp software V. 2.1 (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 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.

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 (17) 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.

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

3.3 Subjects

Sixty-one volunteers completed the experiment, all native speakers of German. The subjects were recruited over the Internet by an announcement on the Language Experiments Portal web page[1] and by postings to relevant newsgroups and mailing lists.

4 Modeling the Judgments

We implemented the methods for selectional preference acquisition which were outlined in Section 2.[2] The algorithms' input were the triple data that had already been extracted for the selection of the experiment's materials (cf. Section 3.1.1).

[1] `http://www.language-experiments.org/`

[2] For Clark and Weir's algorithm, we adapted an existing implementation by Frank Keller and Mirella Lapata from WordNet to GermaNet. We are grateful for their permission to let us use their source code as a basis.

The implementation uses 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.

4.1 Adaptations for Use with GermaNet

For the implementation of Li and Abe's (1998) method, certain modifications to the original GermaNet hierarchy are required. The algorithm operates on a tree, but the GermaNet noun hierarchy is a directed acyclic graph (DAG). As suggested by Li and Abe, each subgraph having multiple parents is copied to transform the DAG into a tree.

A further modification is necessary because in GermaNet, nouns do not only occur as leaves of the hierarchy, but also at internal nodes. Following Wagner (2000) and McCarthy (2001), a new leaf is created for each internal node, containing a copy of the internal node's nouns. This guarantees that all nouns are present at the leaf level.

Finally, the algorithm requires a single root node for the hierarchy. For WordNet and GermaNet, an artificial concept ⟨root⟩ has to be created and connected to the existing top-level classes. WordNet (Version 1.7) incorporates 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 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, we set the number of classes below ⟨root⟩ as a parameter (see Section 4.2).

4.2 Parameter Settings

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 1. For selectional association, the choice was between the highest value, as suggested by Resnik, and the mean value over all classes.

SelA	TCM		SimC			
			highest		mean	
	highest	mean	G^2	χ^2	G^2	χ^2
highest/ mean	125/49/40/33 classes below ⟨root⟩		$\alpha = .0005/.05/.3/.75/.995$			

Table 1: Explored parameter settings

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. Furthermore, as described in Section 4.1, we varied the number of classes below the artificial concept ⟨root⟩. We excluded from the hierarchy classes with less than or equal to 10, 20, and 30 hyponym classes. This resulted in 49, 40, and 33 classes below ⟨root⟩. We also experimented with the 125 classes having at least one hyponym.

Finally, for Clark and Weir's approach, there was a choice between highest and mean value when a noun was ambiguous, between χ^2 and G^2 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).

5 Results

The human judgment data were first normalized by dividing each numerical judgment by the modulus value which 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. The computed preference values were also log-transformed.

Correlation analyses were performed to assess the degree of linear relationship between the human judgments as the dependent variable and the algorithms' selectional preferences with each of the possible parameter settings, corresponding to 30 different independent variables. We examined the subject, direct object, and PP object sentences in isolation as well as at the 90 sentences altogether. Table 2 lists the best correlation coefficients per preference measure, indicating the respective parameters where

Rating	Freq	CondP	SelA	TCM	SimC
SUBJ	.386*	.010	.**408***	.281	.268
			[highest]	[mean, 40 c.b.r.]	[mean, G^2, $\alpha = .75$]
OBJ	.360	.399*	.430*	.251	.**611*****
			[mean]	[mean, 40 c.b.r.]	[highest, G^2, $\alpha = .05$]
PP-OBJ	.168	.335	.330	.319	.**597*****
			[mean]	[mean, 33 c.b.r.]	[highest, G^2, $\alpha = .3$]
overall	.301**	.**374*****	.**374*****	.341***	.232*
			[highest]	[mean, 40 c.b.r.]	[highest, G^2, $\alpha = .3$]

$^{*}p \leq .05$ $^{**}p \leq .01$ $^{***}p \leq .001$ c.b.r.: classes below ⟨root⟩

Table 2: Best correlations between human ratings and selectional preference models

appropriate. For each grammatical relation, the optimal coefficient is emphasized.

The preference measures performed differently well for the three grammatical relations in question. Selectional association (SelA) is the best to model judgments on subjects, closely followed by the simple frequency measure (Freq). The similarity-class method (SimC) yields middle correlations for the direct object relation as well as for the PP object relation.

The highest overall correlation is revealed by conditional probability (CondP) and SelA, closely followed by the tree cut models (TCM). CondP as expressed in (2) outperformed (1), and therefore the latter was excluded from further comparisons.

Regarding the parameter settings, TCM seems to work best with the mean preference value and 40 classes below ⟨root⟩, and SimC yields optimal results using the highest value and the G^2 statistic.

6 Discussion

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.

Although all measures are positively correlated with the human judgments, several of them do not reach significance. Especially, CondP cannot predict subject preferences, and the frequency measure is not suitable for PP objects.

SimC is clearly the optimal predictor for direct objects and PP objects. The tie between verbs and their subjects is less strong than that between verbs and the other arguments; therefore, selectional preferences for subjects are harder to model, which is reflected in the results.

TCM does not reach significant results for any of the individual grammatical relations. This might be due to the fact that the GermaNet noun hierarchy, incorporating 23,053 noun synsets, is considerably smaller than the one of WordNet (version 1.7), which includes 74,488 synsets, and is reduced even more by the parameter for classes below ⟨root⟩.

Overall, the results indicate that the explored approaches to selectional preference acquisition are indeed valid for a language other than English and that they work for grammatical relations other than direct object.

7 Further Research

The correlations found in the evaluation of the approaches to selectional preference acquisition were reasonably high, but there is still room for improvement. The results show that different methods are suited for different argument relations. Therefore, it seems promising to explore model combination using multiple regression to obtain a better fit with the experimental data.

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, e.g., the statistical grammar model described by Schulte im Walde et al. (2001).

Finally, we plan to consider other approaches to selectional preference acquistion: Abney and Light's (1999) hidden Markov models as well as Ciaramita and Johnson's (2000) Bayesian belief networks. We also intend to vary the size of the taxonomy and explore the effects on the computed selectional preference values.

References

Steven Abney and Marc Light. 1999. Hiding a semantic hierarchy in a Markov model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing at the 37th ACL*, pages 1–8, University of Maryland, College Park, MD, USA.

Daniel M. Bikel. 2000. A statistical model for parsing and word-sense disambiguation. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pages 155–163, Hong Kong.

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

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 187–193, Saarbrücken, Germany.

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

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

Birgit Hamp and Helmut 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 ACL and the 8th EACL*, pages 9–15, Madrid, Spain.

Frank Keller, Martin Corley, Steffan Corley, Lars Konieczny, and Amalia Todirascu. 1998. WebExp: A Java toolbox for web-based psychological experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh, UK.

Maria Lapata, Frank Keller, and Scott 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)*, pages 346–353, Toulouse, France.

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

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

Diana McCarthy, John Carroll, and Judita Preiss. 2001. Disambiguating noun and verb senses using automatically acquired selectional preferences. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2) at the 39th ACL and the 10th EACL*, Toulouse, France.

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

Günter Neumann, Rolf Backofen, Judith Baur, Markus Becker, and Christian 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)*, pages 209–216, Washington, DC, USA.

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

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pages 52–57, Washington, DC, USA.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14:465–471.

Sabine Schulte im Walde, Helmut Schmid, Mats Rooth, Stefan Riezler, and Detlef Prescher. 2001. Statistical grammar models and lexicon acquisition. In Christian Rohrer, Antje Roßdeutscher, and Hans Kamp, editors, *Linguistic Form and its Computation*, pages 389–440. CSLI Publications, Stanford, CA, USA.

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

Andreas Wagner. 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 ECAI*, pages 37–42, Berlin, Germany.