# Evaluating and Combining Approaches to Selectional Preference Acquisition

**Carsten Brockmann**
School of Informatics
The University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
Carsten.Brockmann@ed.ac.uk

**Mirella Lapata**
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK
mlap@dcs.shef.ac.uk

## Abstract

Previous work on the induction of selectional preferences has been mainly carried out for English and has concentrated almost exclusively on verbs and their direct objects. In this paper, we focus on class-based models of selectional preferences for German verbs and take into account not only direct objects, but also subjects and prepositional complements. We evaluate model performance against human judgments and show that there is no single method that overall performs best. We explore a variety of parametrizations for our models and demonstrate that model combination enhances agreement with human ratings.

## 1 Introduction

Selectional preferences or constraints are the semantic restrictions that a word imposes on the environment in which it occurs. A verb like *eat* typically takes animate entities as its subject and edible entities as its object. Selectional preferences can most easily be observed in situations where they are violated. For example, in the sentence "*The mountain eats sincerity.*" both subject and object preferences for the verb *eat* are violated. The problem of quantifying the degree to which a given predicate (e.g., *eat*) semantically fits its arguments has received a lot of attention within computational linguistics. Several approaches have been developed for the induction of selectional preferences, and almost all of them rely on the availability of large machine-readable corpora.

Probably the most primitive corpus-based model of selectional preferences is co-occurrence frequency. Inspection in a corpus of the types of nouns *eat* admits as its objects will reveal that *food, meal, meat,* or *lunch* are frequent complements, whereas *river, mountain,* or *moon* are rather unlikely. The obvious disadvantage of the frequency-based approach is that no generalizations emerge with respect to the observed preferences as it embodies no notion of semantic relatedness or proximity. Ideally, one would like to infer from the corpus that *eat* is semantically congruent with food-related objects and incongruent with natural objects. Another related limitation of the frequency-based account is that it cannot make any predictions for words that never occurred in the corpus. A zero co-occurrence count might be due to insufficient evidence or might reflect the fact that a given word combination is inherently implausible.

For the above reasons, most approaches model the selectional preferences of predicates (e.g., verbs, nouns, adjectives) by combining observed frequencies with knowledge about the semantic classes of their arguments. The classes can be induced directly from the corpus (Pereira et al., 1993; Brown et al., 1992; Lapata et al., 2001) or taken from a manually crafted taxonomy (Resnik, 1993; Li and Abe, 1998; Clark and Weir, 2002; Ciaramita and Johnson, 2000; Abney and Light, 1999). In the latter case the taxonomy is used to provide a mapping from words to conceptual classes, and in most cases WordNet (Miller et al., 1990) is employed for this purpose.

Although most approaches agree on how selectional preferences must be *represented*, i.e., as a mapping $\sigma : (p, r, c) \rightarrow a$ that maps each predicate $p$ and the semantic class $c$ of its argument with respect to role $r$ to a real number $a$ (Light and Greiff, 2002), there is little agreement on how selectional preferences must be *modeled* (e.g., whether to use a probability model or not) and *evaluated* (e.g., whether to use a task-based evaluation or not). Furthermore, previous work has almost exclusively focused on verbal selectional

preferences in English with the exception of Lapata et al. (1999, 2001), who look at adjective-noun combinations, again for English. Verbs tend to impose stricter selectional preferences on their arguments than adjectives or nouns and thus provide a natural test bed for models of selectional preferences. However, research on verbal selectional preferences has been relatively narrow in scope as it has primarily focused on verbs and their direct objects, ignoring the selectional preferences pertaining to subjects and prepositional complements.

The induction of selectional preferences typically addresses two related problems: (a) finding an appropriate class that best fits the predicate in question and (b) coming up with a statistical model or a measure that estimates how well a predicate fits its arguments. Resnik (1993) defines *selectional association*, an information-theoretic measure of semantic fit of a particular semantic class $c$ as an argument to a predicate $p$. Li and Abe (1998) use the Minimum Description Length (MDL) principle to select the the appropriate class $c$, Clark and Weir (2002) employ hypothesis testing. Abney and Light (1999) propose Hidden Markov Models as a way of deriving selectional preferences over words, senses, or even classes, whereas Ciaramita and Johnson (2000) use Bayesian Belief Networks to quantify selectional preferences.

Although there is no standard way to evaluate different approaches to selectional preferences, two types of evaluation are usually conducted: task-based evaluation and comparisons against human judgments. Word sense disambiguation results are reported by Resnik (1997), Abney and Light (1999), Ciaramita and Johnson (2000) and Carroll and McCarthy (2000) (however, on a different data set). Among the first three approaches, Ciaramita and Johnson (2000) obtain the best results. Li and Abe (1998) evaluate their system on the task of prepositional phrase attachment, whereas Clark and Weir (2002) use pseudo-disambiguation,[1] a somewhat artificial task, and show that their approach outperforms Li and Abe (1998) and Resnik (1993).

Another way to evaluate a model's performance is agreement with human ratings. This can be done by selecting predicate-argument structures randomly, using the model to predict the degree of semantic fit and then looking at how well the ratings

---

[1]The task is to decide which of two verbs $v_1$ and $v_2$ is more likely to take a noun $n$ as its object. The method being tested must reconstruct which of the unseen $(v_1, n)$ and $(v_2, n)$ is a valid verb-object combination.

correlate with the model's predictions (Resnik, 1993; Lapata et al., 1999; Lapata et al., 2001). This approach seems more appropriate for languages for which annotated corpora with word senses are not available. It is more direct than 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. It is also more natural than pseudo-disambiguation which relies on artificially constructed data sets. Large-scale comparative studies have not, however, assessed the strengths and weaknesses of the proposed methods as far as modeling human data is concerned.

In this paper, we undertake such a comparative study by looking at selectional preferences of German verbs. In contrast to previous work, we take into account not only verbs and their direct objects, but also subjects and prepositional complements. We focus on three previously well-studied models, Resnik's (1993) selectional association, Li and Abe's (1998) MDL and Clark and Weir's (2002) probability estimation method. For comparison, we also employ two models that do not incorporate any notion of semantic class, namely co-occurrence frequency and conditional probability.

In the remainder of this paper, we briefly review the models of selectional preferences we consider (Section 2). Section 3 details our experiments, evaluation methodology, and reports our results. Section 4 offers some discussion and concluding remarks.

## 2 Models of Selectional Preferences

**Co-occurrence Frequency.** We can quantify the semantic fit between a verb and its arguments by simply counting $f(v, r, n)$, the number of times a noun $n$ co-occurs with a verb $v$ in a grammatical relation $r$.

**Conditional Probability.** As we discuss below, most class-based approaches to selectional preferences rely on the estimation of the conditional probability $P(n|v, r)$, where $n$ is represented by its corresponding classes in the taxonomy. Here we concentrate solely on the nouns as attested in the corpus without making reference to a taxonomy and estimate the following:

$$\hat{P}(n|v, r) = \frac{f(v, r, n)}{f(v, r)} \qquad (1)$$

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

In (1) it is the verb that imposes the semantic preferences on its arguments, whereas in (2) selectional preferences are expressed in the other direction, i.e. arguments select for their predicates.

**Selectional Association.** Resnik (1993) was the first to propose a measure of the the semantic fit of a particular semantic class $c$ as an argument to a verb $v$. *Selectional association* (see (3) and (4)) represents the contribution of a particular semantic class $c$ to the total quantity of information provided by a verb about the semantic classes of its argument, when measured as the relative entropy between the prior distribution of classes $P(c)$ and the posterior distribution $P(c|v,r)$ of the argument classes for a particular verb $v$. The latter distribution is estimated as shown in (5).

$$A(v,r,c) = \frac{P(c|v,r)\log\frac{P(c|v,r)}{P(c)}}{\eta} \qquad (3)$$

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

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

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):

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

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

**Tree Cut Models.** Li and Abe (1998) use MDL to select from a hierarchy a set of classes that represent the selectional preferences for a given verb. These preferences are probabilities of the form $P(n|v,r)$ where $n$ is a noun represented by a class in the taxonomy, $v$ is a verb and $r$ is an argument slot. Li and Abe's algorithm operates on thesaurus-like hierarchies where each leaf node stands for a noun, each internal node stands for the class of nouns below it, and a noun is uniquely represented by a leaf node. Li and Abe derive a separate model for each verb by partitioning the leaf nodes (i.e., nouns) of the thesaurus tree and associating a probability with each class in the partition.

More formally, a *tree cut model* $M$ is defined as a pair of a tree cut $\Gamma$, which is a set of classes $c_1, c_2, \ldots, c_k$, and a parameter vector $\theta$ specifying a probability distribution over the members of $\Gamma$ with the constraint that the probabilities sum to one.

$$\sum_{i=1}^{k} P(c_i|v,r) = 1 \qquad (7)$$

To select the tree cut model that best fits the data, Li and Abe (1998) employ the MDL principle (Rissanen, 1978) by considering the cost in bits of describing both the model itself and the observed data (in our case verb-argument combinations).

Given a data sample $S$ encoded by a tree cut model $\hat{M} = (\Gamma, \hat{\theta})$ with tree cut $\Gamma$ and estimated parameters $\hat{\theta}$, the total description length in bits $L(\hat{M}, S)$ is given by equation (8):

$$L(\hat{M}, S) = \log|G| + \frac{k}{2}\log|S| \\ - \sum_{n\in S}\log P_{\hat{M}}(n|v,r) \qquad (8)$$

where $|G|$ is the cardinality of the set of all possible tree cuts, $k$ is the number of classes on the cut $\Gamma$, $|S|$ is the sample size, and $P_{\hat{M}}(n|v,r)$ is the probability of a noun, which is estimated by distributing the probability of a given class equally among the nouns that can be denoted by it:

$$\forall n\in\mathrm{syn}(c): P_{\hat{M}}(n|v,r) = \frac{P_{\hat{M}}(c|v,r)}{|\mathrm{syn}(c)|} \qquad (9)$$

**Class-based Probability.** Clark and Weir (2002) are, strictly speaking, not concerned with the induction of selectional preferences but with the problem of estimating conditional probabilities of the form shown in (1) in the face of sparse data. However, their probability estimation method can be naturally applied to the selectional preference acquisition problem as it is suited not only for the estimation of the appropriate probabilities but also for finding a suitable class for the predicates of interest. Clark

and Weir obtain the probability $P(v|c,r)$ from $P(c|v,r)$ using Bayes' theorem:

$$P(c|v,r) = P(v|c,r)\frac{P(c|r)}{P(v|r)} \qquad (10)$$

They suggest the following way for finding a set of concepts $\overline{c'}$ (where $\overline{c'}$ denotes the set of concepts dominated by $c'$, including $c'$ itself) as a generalization for concept $c$ (where $c$ can be either $n$ or one of its hypernyms): Initially, $c'$ is set to $c$, then $c'$ is set to successive hypernyms of $c$ until a node in the hierarchy is reached where $P(\overline{c'}|v,r)$ changes significantly. This is determined by comparing estimates of $P(\overline{c'_i}|v,r)$ for each child $c'_i$ of $c'$ using hypothesis testing. 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 a significant difference between them, the null hypothesis is rejected and classes that are lower in the hierarchy than $c'$ are used. Selecting the right level of generalization crucially depends on the type of statistic used (in their experiments Clark and Weir use the Pearson chi-square statistic $\chi^2$ and the log-likelihood chi-square statistic $G^2$). The appropriate level of significance $\alpha$ can be tuned experimentally.

Once a suitable class is found, the *similarity-class probability* $P_{sc}$ is estimated:

$$P_{sc}(c|v,r) = \frac{\hat{P}(v|[v,r,c],r)\frac{\hat{P}(c|r)}{\hat{P}(v|r)}}{\sum_{c'\in C}\hat{P}(v|[v,r,c'],r)\frac{\hat{P}(c'|r)}{\hat{P}(v|r)}} \qquad (11)$$

where $[v,r,c]$ 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. The denominator is a normalization factor. 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 (5)).

## 3 Experiments

### 3.1 Parameter Settings

In our experiments, we compared the performance of the five methods discussed above against human judgments. Before discussing the details of our evaluation we present our general experimental setup (e.g., the corpora and hierarchy used) and the different types of parameters we explored.

All our experiments were conducted on data obtained from the German Süddeutsche Zeitung (SZ)

corpus, a 179 million word collection of newspaper texts. The corpus was parsed using the grammatical relation recognition component of SMES, a robust information extraction core system for the processing of German text (Neumann et al., 1997). SMES incorporates a tokenizer that maps the text into a stream of tokens. The tokens are then analyzed morphologically (compound recognition, assignment of part-of-speech tags), and a chunk parser identifies phrases and clauses by means of finite state grammars. The grammatical relations recognizer operates on the output of the parser while exploiting a large subcategorization lexicon. Although SMES recognizes a variety of grammatical relations, in our experiments we focused solely on relations of the form $(v,r,n)$ where $r$ can be a subject, direct object, or prepositional object (see the examples in Table 2).

For the class-based models, the hierarchy available in GermaNet (Hamp and Feldweg, 1997) was used. The experiments reported in this paper make use of the noun taxonomy of GermaNet (version 3.0, 23,053 noun synsets), and the information encoded in it in terms of the hyponymy/hypernymy relation.

Certain modifications to the original GermaNet hierarchy were necessary for the implementation of Li and Abe's method (1998). The GermaNet noun hierarchy is a directed acyclic graph (DAG) whereas their algorithm operates on trees. A solution to this problem is given by Li and Abe, who transform the DAG into a tree by copying each subgraph having multiple parents. An additional modification is needed since in GermaNet, nouns do not only occur as leaves of the hierarchy, but also at internal nodes. Following Wagner (2000) and McCarthy (2001), we 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.

Finally, the algorithm requires that the employed hierarchy has a single root node. In WordNet and GermaNet, nouns are not contained in a single hierarchy; instead they are partitioned according to a set of semantic primitives which are treated as the unique beginners of separate hierarchies. This means that an artificial concept $\langle\text{root}\rangle$ has to be created and connected to the existing top-level classes. Although WordNet has only nine classes without a hypernym, GermaNet contains 502. Of these, 125 have one or more daughters.

The number of classes below $\langle\text{root}\rangle$ has an immediate effect on the tree cut model: With a large

| SelA | TCM | | SimC | | | |
| --- | --- | --- | --- | --- | --- | --- |
| | | | highest | | mean | |
| | highest | mean | $G^2$ | $\chi^2$ | $G^2$ | $\chi^2$ |
| highest, mean | 33 c.b.r., 49 c.b.r., | 40 c.b.r., 125 c.b.r. | $\alpha = .0005$, $\alpha = .3$, | $\alpha = .05$, $\alpha = .75$, $\alpha = .995$ | | |

c.b.r.: classes below ⟨root⟩

Table 1: Explored parameter settings

number of classes, many of the cuts returned by MDL are over-generalizing at the ⟨root⟩ level. We therefore varied the the number of classes below ⟨root⟩ in order to observe how this affects the generalization outcome. We excluded from the hierarchy classes with less than or equal to 10, 20, and 30 hyponyms. This resulted in 49, 40, and 33 classes below ⟨root⟩. We also experimented with the full 125 classes (see Table 1).

All of the class-based methods produce a value for each class $c$ to which an argument noun $n$ belongs. Since $n$ can be ambiguous and its appropriate sense is not known, a unique class is typically chosen by simply selecting the class which maximizes the quantity of interest (see (3), (9), and (11)). An alternative is to consider the mean value over all classes. In our experiments, we compare the effect of these distinct selection procedures.

Finally, for Clark and Weir's (2002) approach, two parameters are important for finding an appropriate generalization class: (a) the statistic for performing significance testing and (b) the $\alpha$ value for determining the significance level. Here, we experimented with the $\chi^2$ and $G^2$ statistics and ran our experiments for the following different $\alpha$ values: .0005, .05, .3, .75, and .995. The parameter settings we explored are shown in Table 1.

## 3.2 Eliciting Judgments on Selectional Preferences

In order to evaluate the methods introduced in Section 2, we first established an independent measure of how well a verb fits its arguments by eliciting judgments from human subjects (Resnik, 1993; Lapata et al., 2001; Lapata et al., 1999). In this section, we describe our method for assembling the set of experimental materials and collecting plausibility ratings for these stimuli.

**Materials and Design.** As mentioned earlier, co-occurrence triples of the form $(v, r, n)$ were extracted from the output of SMES. 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 fre-

quency less than one per million. Ten verbs were selected randomly for each grammatical relation. For each verb we divided the set of triples into three bands (High, Medium, and Low), based on an equal division of the range of log-transformed co-occurrence frequency, and randomly chose one noun from each band. The division ensured that the experimental stimuli represented likely and unlikely verb-argument combinations and enabled us to investigate how the different models perform with low/high counts. Example stimuli are shown in Table 2.

Our 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 10 common human first names (five female, five male) was added as subject where possible, or else an inanimate subject which appeared frequently in the corpus was chosen.

**Procedure.** The experimental paradigm was Magnitude Estimation (ME), a technique standardly used in psychophysics to measure judgments of sensory stimuli (Stevens, 1975), which Bard et al. (1996) and Cowart (1997) have applied to the elicitation of linguistic judgments. ME has been shown to provide fine-grained measurements of linguistic acceptability which are robust enough to yield statistically significant results, while being highly replicable both within and across speakers.

ME requires subjects to assign numbers to a series of linguistic stimuli in a proportional fashion. Subjects are first exposed to a modulus item, to which they assign an arbitrary number. All other stimuli are rated proportionally to the modulus. In this way, each subject can establish their own rating scale.

In the present experiment, the subjects were instructed to judge how acceptable the 90 sentences were in proportion to a modulus sentence. The experiment was conducted remotely over the Internet using WebExp 2.1 (Keller et al., 1998), an interactive software package for administering web-based psychological experiments. Subjects first saw a set of instructions that explained the ME technique and included some examples, and had to fill in a short questionnaire including basic demographic information. Each subject saw 90 experimental stimuli. A random stimulus order was generated for each subject.

| Relation | Verb | Co-occurrence Frequency Band | | | | | |
|---|---|---|---|---|---|---|---|
| | | High | | Medium | | Low | |
| SUBJ | stagnieren stagnate | Umsatz turnover | 1.77 | Preis price | .85 | Arbeitslosigkeit unemployment | .48 |
| OBJ | erlegen shoot | Tier animal | .60 | Jahr year | .30 | Gesetz law | 0 |
| PP-OBJ | denken an think of | Rücktritt resignation | 1.54 | Freund friend | .78 | Kleinigkeit detail | 0 |

Table 2: Example stimuli (with log co-occurrence frequencies in the SZ corpus)

| Rating | ISAgr | Freq | CondP | SelA | TCM | SimC |
|---|---|---|---|---|---|---|
| SUBJ | .790 | .386* | .010 | **.408*** [highest] | .281 [mean, 40 c.b.r.] | .268 [mean, $G^2$, $\alpha = .75$] |
| OBJ | .810 | .360 | .399* | .430* [mean] | .251 [mean, 40 c.b.r.] | **.611**\*** [highest, $G^2$, $\alpha = .05$] |
| PP-OBJ | .820 | .168 | .335 | .330 [mean] | .319 [mean, 33 c.b.r.] | **.597**\*** [highest, $G^2$, $\alpha = .3$] |
| overall | .810 | .301** | **.374**\*** | **.374**\*** [highest] | .341*** [mean, 40 c.b.r.] | .232* [highest, $G^2$, $\alpha = .3$] |

$^*p \leq .05$ $^{**}p \leq .01$ $^{***}p \leq .001$ c.b.r.: classes below $\langle \texttt{root} \rangle$

Table 3: Best correlations between human ratings and selectional preference models

**Subjects.** The experiment was completed by 61 volunteers, all self-reported native speakers of German. Subjects were recruited via postings to Usenet newsgroups.

### 3.3 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 (Bard et al., 1996). All analyses were conducted on the normalized, log-transformed judgments.

Using correlation analysis we explored the linear relationship between the human judgments and the methods discussed in Section 2. As shown in Table 1 there are 30 distinct parameter instantiations for the class-based models. There are no parameters for co-occurrence frequency and conditional probability. Table 3 lists the best correlation coefficients per method, indicating the respective parameters where appropriate. For each grammatical relation, the optimal coefficient is emphasized.

In Table 3, we also show how well humans agree in their judgments (inter-subject agreement, ISAgr) and thus provide an upper bound for

the task which allows us to interpret how well the models are doing in relation to humans. We performed correlations on the elicited judgments using leave-one-out resampling (Weiss and Kulikowski, 1991). We divided the set of the subjects' responses with size $m$ into a set of size $m - 1$ (i.e., the response data of all but one subject) and a set of size one (i.e., the response data of a single subject). We then correlated the mean rating of the former set with the rating of the latter. This was repeated $m$ times and the average agreement is reported in Table 3.

As shown in Table 3, all five models are significantly correlated with the human ratings, although the correlation coefficients are not as high as the inter-subject agreement (ISAgr). Selectional association (SelA) and conditional probability (CondP) reveal the highest overall correlations. CondP as expressed in (2) outperformed (1) which was excluded from further comparisons. As far as the individual argument relations are concerned, the similarity-class probability (SimC) performs best at modeling the selectional preferences for prepositional and direct objects. Clark and Weir's (2002) pseudo-disambiguation experiments also show that their method outperforms tree cut models (TCM) and SelA at modeling the semantic fit between verbs and their direct objects. Our results additionally generalize to PP-objects. SelA is the best predictor for subject-related selectional pref-

| Factor | Eigenvalue | Variance | Cumulative |
|--------|-----------|----------|------------|
| SimC   | 7.969     | 53.1%    | 53.1%      |
| TCM    | 3.251     | 21.7%    | 74.8%      |
| SelA   | 1.185     | 7.9%     | 82.7%      |
| CondP  | 0.853     | 5.7%     | 88.4%      |

Table 4: Principal component factors

erences, whereas co-occurrence frequency (Freq) is the second best.

With respect to the class selection method, better results are obtained when the highest class is chosen. This is true for SelA and SimC but not for TCM where the mean generally yields better performance. Recall from Section 3.1 that for TCM the number of classes below ⟨root⟩ was varied from 125 to 33. As can be seen from Table 3, better results are obtained with 40 and 33 classes, i.e., with a relatively small number of classes below ⟨root⟩. Finally, in agreement with Clark and Weir, for SimC the best results were obtained with the $G^2$ statistic. Also note that different $\alpha$ values seem to be appropriate for different argument relations.

### 3.4 Model Combination

An obvious question is whether a better fit with the experimental data can be obtained via model combination. As discussed earlier different models seem to provide complementary information when it comes to modeling different argument relations. A straightforward way to combine our different models is multiple linear regression. Recall that we have 30 variants of class-based models (only the best performing ones are shown in Table 3), some of which are expectedly highly correlated. After removing models with high intercorrelation ($r \geq .99$, 15 out of 30), principal components factor analysis (PCFA) was performed on all 90 items, keeping the factors that explained more than 5% of the variance (see Table 4).

Multiple regression on all 90 observations with all four factors and forward selection (with $p \geq .05$ for removal from the model) yielded the regression equation in (12). The corresponding correlation coefficient is .47 ($p \leq .001$).

$$\text{Rating} = .091 \text{ CondP} + .068 \text{ TCM} \\ + .103 \text{ SelA} + .052 \tag{12}$$

Equation (12) was derived from the entire data set (i.e., 90 verb-argument combinations). Ideally, one would need to conduct another experiment with a new set of materials in order to determine whether (12) generalizes to unseen data. In default

of a second experiment which we plan for the future, we investigated how well model combination performs on unseen data by using 10-fold cross-validation.

Our data set was split into 10 disjoint subsets each containing 9 items. We repeated the PCFA procedure and the multiple regression analysis 10 times, each time using 81 items as training data and the remaining 9 as test data. Then we performed a correlation analysis between the predicted values for the unseen items of each fold and the human ratings. Effectively, this analysis treats the whole data set as unseen. However notice that for each test/train set split we obtain different regression equations since the PCFA yields different factors for different data sets. Comparison between the estimated values and the human ratings yielded a correlation coefficient of .40 ($p \leq .001$) outperforming any single model.

### 4 Discussion

In this paper, we evaluated five models for the acquisition of selectional preferences. We focused on German verbs and their subjects, direct objects, and PP-objects. We placed emphasis on class-based models of selectional preferences, explored their parameter space, and showed that the existing models, developed primarily for English, also generalize to German. We proposed to evaluate the different models against human ratings and argued that such an evaluation methodology allows us to assess the feasibility of the task and to compute performance upper bounds.

Our results indicate that there is no method which overall performs best; it seems that different methods are suited for different argument relations (i.e., SimC for objects, SelA for subjects). The more sophisticated class-based approaches do not always yield better results when compared to simple frequency-based models. This is in agreement with Lapata et al. (1999) who found that co-occurrence frequency is the best predictor of the plausibility of adjective-noun pairs. Model combination seems promising in that a better fit with experimental data is obtained. However, note that none of our models (including the ones obtained via multiple regression) seem to attain results reasonably close to the upper bound.

In the future, we plan to consider web-based frequencies for our probability estimates (Keller et al., 2002) as well as Abney and Light's (1999) Hidden Markov Models and Ciaramita and Johnson's (2000) Bayesian Belief Networks. We will also expand our evaluation methodol-

ogy to adjective-noun and noun-noun combinations and conduct further rating experiments to cross-validate our combined models.

# References

Steve Abney and Marc Light. 1999. Hiding a semantic class hierarchy in a Markov model. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8, College Park, MD.

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

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

John Carroll and Diana McCarthy. 2000. Word sense disambiguation using automatically acquired verbal preferences. *Computers and the Humanities*, 34(1–2):109–114.

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional restrictions with Bayesian networks. In *Proceedings of the 18th International Conference on Computational Linguistics*, 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.

Wayne Cowart. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Sage Publications, Thousand Oaks, CA.

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.

Frank Keller, Maria Lapata, and Olga Ourioupina. 2002. Using the web to overcome data sparseness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Philadelphia, PA.

Maria Lapata, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 30–36, Bergen, Norway.

Maria Lapata, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgments. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 346–353, Toulouse, France.

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

Marc Light and Warren Greiff. 2002. Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 87:1–13.

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

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

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 ACL Conference on Applied Natural Language Processing*, pages 209–216, Washington, DC.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.

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

Philip Resnik. 1997. Selectional preferences 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.

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

S. S. Stevens. 1975. *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects*. John Wiley & Sons, New York, NY.

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 at the 14th ECAI*, pages 37–42, Berlin, Germany.

Sholom M. Weiss and Casimir A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.