# A Comparison of Selectional Preference Models for Automatic Verb Classification

**Will Roberts** and **Markus Egg**
Institut für Anglistik und Amerikanistik, Humboldt University
10099 Berlin, Germany
{will.roberts,markus.egg}@anglistik.hu-berlin.de

## Abstract

We present a comparison of different selectional preference models and evaluate them on an automatic verb classification task in German. We find that all the models we compare are effective for verb clustering; the best-performing model uses syntactic information to induce nouns classes from unlabelled data in an unsupervised manner. A very simple model based on lexical preferences is also found to perform well.

## 1 Introduction

Selectional preferences (Katz and Fodor, 1963; Wilks, 1975; Resnik, 1993) are the tendency for a word to semantically select or constrain which other words may appear in a direct syntactic relation with it. Selectional preferences (SPs) have been a perennial knowledge source for NLP tasks such as word sense disambiguation (Resnik, 1997; Stevenson and Wilks, 2001; McCarthy and Carroll, 2003) and semantic role labelling (Erk, 2007); and recognising selectional violations is thought to play a role in identifying and interpreting metaphor (Wilks, 1978; Shutova et al., 2013). We focus on the SPs of verbs, since determining which arguments are typical of a given verb sheds light on the semantics of that verb.

In this study, we present the first empirical comparison of different SP models from the perspective of automatic verb classification (Schulte im Walde, 2009; Sun, 2012), the task of grouping verbs together based on shared syntactic and semantic properties.

We cluster German verbs using features capturing their *valency* or *subcategorisation*, following prior work (Schulte im Walde, 2000; Esteve Ferrer, 2004; Schulte im Walde, 2006; Sun et al., 2008; Korhonen et al., 2008; Li and Brew, 2008), and investigate the effect of adding information about verb argument preferences. SPs are represented by features capturing lexical information about the heads of arguments to the verbs; we restrict our focus here to nouns.

We operationalise a selectional preference model as a function which maps such an argument head to a concept label. We submit that the primary characteristic of such a model is its *granularity*. In our baseline condition, all nouns are mapped to the same label; this effectively captures no information about a verb's SPs (i.e., we cluster verbs using subcategorisation information only). On the other extreme, each noun is its own concept label; we term this condition *lexical preferences* (LP). Between the baseline and LP lie a spectrum of models, in which multiple concepts are distinguished, and each concept label can represent multiple nouns. Our main hypothesis is that verb clustering will work best using a model of such intermediate granularity. This follows the intuition that verbs would seem to select for classes of nouns; for instance, we suppose that *essen* 'eat' would tend to prefer as a direct object a noun from the abstract concept *Essen* ('food'). We assume that these concepts can be expressed independently of particular predicates; that is, there exist selectional preference models that will work for all verbs (and all grammatical relations). Further benefits of grouping nouns into classes include combating data sparsity, as well as deriving models which can generalise to nouns unseen in training data.

Another parameter of a selectional preference model is the methodology used to induce the conceptual classes; put another way, the success of an SP model hinges on how it represents concepts. In this paper, we investigate the choice of noun categorisation method through an empirical comparison of selectional preference models previously used in the literature.

We set out to investigate the following questions:

1. What classes of nouns are effective descriptors



*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 511–522,
October 25-29, 2014, Doha, Qatar. 

of selectional preference concepts? For example, do they correspond to features such as ANIMATE?

2. What is the appropriate granularity of selectional preference concepts?

3. Which methods of classifying nouns into concepts are most effective at capturing selectional preferences for verb clustering?

This paper is structured as follows: In Section 2, we introduce our baseline method of clustering verbs using subcategorisation information and describe evaluation; Section 3 lists the models of selectional preferences that we compare in this work; Section 4 presents results and discussion; Section 5 summarises related work; and Section 6 concludes with directions for future research.

## 2 Automatic verb classification

Verb classifications such as VerbNet (Kipper-Schuler, 2005) allow generalisations about the syntax and semantics of verbs and have proven useful for a range of NLP tasks; however, creation of these resources is expensive and time-consuming. Automatic verb classification seeks to learn verb classes automatically from corpus data in a cheaper and faster way. This endeavour is possible due to the link between a verb's semantics and its syntactic behaviour (Levin, 1993). Recent research has found that even automatically-acquired classifications can be useful for NLP applications (Shutova et al., 2010; Guo et al., 2011). In this section, we introduce the verb classification method used by our baseline model, which clusters verbs based on subcategorisation information. Following this, Section 2.2 explains the gold standard verb clustering and cluster purity metric which we use for evaluation.

### 2.1 Baseline model

In this work, we take subcategorisation to mean the requirement of a verb for particular types of argument or concomitant. For example, the English verb *put* subcategorises for subject, direct object, and a prepositional phrase (PP) like *on the shelf*:

(1) *[$_{NP}$ Al] put [$_{NP}$ the book] [$_{PP}$ on the shelf].*

A *subcategorisation frame* (SCF) describes a combination of arguments required by a specific verb; a description of the set of SCFs which a verb may take is called its *subcategorisation preference*. We acquire descriptions of verbal SCF preferences on the basis of unannotated corpus data.

Our experiments use the SdeWaC corpus (Faaß and Eckart, 2013), containing 880 million words in 45 million sentences; this is a subset of deWaC (Baroni et al., 2009), a corpus of $10^9$ words extracted from Web search results. SdeWaC is filtered to include only those sentences which are maximally parsable[1]. We parsed SdeWaC with the `mate-tools` dependency parser (Bohnet et al., 2013)[2], which performs joint POS and morphological tagging, as well as lemmatisation. Our subcategorisation analyses are delivered by the rule-based SCF tagger described by Roberts et al. (2014), which operates using the dependency parses and assigns each finite verb an SCF type. The SCF tags are taken from the SCF inventory proposed by Schulte im Walde (2002), which indicates combinations of nominal and verbal complement types, such as `nap:für.Acc` (transitive verb, with a PP headed by *für* 'for'). Examples of complements are `n` for nominative subject, and `a` for accusative direct object; in SCFs which include PPs (`p`), the SCF tag specifies the head of the PP and the case of the prepositional argument (`Acc` in our example indicates the accusative case of the prepositional argument). The SCF tagger undoes passivisation and analyses verbs embedded in modal and tense constructions. We record 673 SCF types in SdeWaC.

From SdeWaC, we extracted the first 3,000,000 verb instances assigned an SCF tag by the SCF tagger, where the verb lemma is one of the 168 listed in our gold standard clustering (this requires approximately 270 million words of parsed text, or 25% of SdeWaC). We refer to this as our **test set**. In this set, each verb is seen on average 17,857 times; the most common is *geben* ('give', 328,952 instances), and the least is *grinsen* ('grin', 50).

We represent verbs as vectors, where each dimension represents a different SCF type. Vector entries are initialised with SCF code counts over the test set, and each vector is then normalised to sum to 1, so that a vector represents a discrete probability distribution over the SCF inventory. We use the Jensen-Shannon divergence as a dissimilarity measure between pairs of verb vectors. The Jensen-Shannon divergence (Lin, 1991) is an information-theoretic, symmetric measure (Equation (2)) re-

[1] The filtering used a rule-based dependency parser to estimate a per-token parse error rate for each sentence, and removed those sentences with very high error rates.

[2] `https://code.google.com/p/mate-tools/`

lated to the Kullback-Leibler divergence (Equation (3)).

$$JS(p,q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2}) \quad (2)$$

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \quad (3)$$

With this dissimilarity measure, we use hierarchical clustering with Ward's criterion (Ward, Jr, 1963) to partition the verbs into $K$ disjoint sets (i.e., hard clustering), where we match $K$ to the number of classes in our gold standard (described below).

### 2.2 Evaluation paradigm

We evaluate the automatically induced verb clusterings against a manually-constructed gold standard, published by Schulte im Walde (2006, page 162ff.). This Levin-style classification groups 168 high- and low-frequency verbs into 43 semantic classes; examples include Aspect (e.g., *anfangen* 'begin'), Propositional Attitude (e.g., *denken* 'think'), and Weather (e.g., *regnen* 'rain'). Some of the classes are further sub-classified; for the purposes of our evaluation, we ignore the hierarchical structure of the classification and consider each class or subclass to be a separate entity. In this way, we obtain classes of fairly comparable size and sufficient semantic consistency.[3]

We evaluate a given verb clustering against the gold standard using the *pairwise $F$-score* (Hatzivassiloglou and McKeown, 1993). To calculate this statistic, we construct a contingency table over the $\binom{n}{2}$ pairs of verbs, the idea being that the gold standard provides binary judgements about whether two verbs should be clustered together or not. If a clustering agrees with the gold standard as to whether a pair of verbs belong together or not, this is a "correct" answer. Using the contingency table, the standard information retrieval measures of precision ($P$) and recall ($R$) can be computed; the $F$-score is then the harmonic mean of these: $F = 2PR/(P+R)$. The random baseline is 2.08 (calculated as the average score of 50 random partitions), and the optimal score is 95.81, calculated by evaluating the gold standard against itself. As the gold standard includes polysemous verbs, which belong to more than one cluster, the optimal score is calculated by randomly picking one of their senses; the average is then taken over 50 such trials.

The pairwise $F$-score is known to be somewhat nonlinear (Schulte im Walde, 2006), penalising early clustering "mistakes" more than later ones, but it has the advantage that we can easily determine statistical significance using the contingency table and McNemar's test.

We use only one clustering algorithm and one purity metric, because our prior work shows that the most important choices for verb clustering are the distance measure used, and how verbs are represented. These factors set, we expect similar performance trends from different algorithms, with predictable variation (e.g., spectral tends to outperform hierarchical clustering, which in turn outperforms $k$-means). Combining Ward's criterion and $F$-score is a trade-off at this point; the criterion is deterministic, giving reproducible results without computational complexity, but disallows estimates of density over our evaluation metric and is greedy (see discussion in Section 4.3).

## 3 Selectional preference models

In this section, we introduce the various SP models that we compare in this paper. In all cases, we hold the verb clustering procedure described in the previous section unchanged, with the exception that SCF tags for verbs are parameterised for selectional preferences. As an example, a verb instance observed in a simple transitive frame with a nominal subject and accusative object would receive the SCF tag `na`. Assuming that a given SP model places the subject noun in the SP concept `animate` and the object noun in the concept `concrete`, the parameterised SCF tag would be `na*subj-{animate}*obj-{concrete}`. This process captures argument co-occurrence information about verb instances, and has the effect of multiplying the SCF inventory size, making the verb vectors described in Section 2.1 both longer and sparser.

We evaluate various types of SP models: the simple lexical preferences model; three models which perform automatic unsupervised induction of noun concepts from unlabelled data; and one which uses a manually-built lexical resource. As far as we are aware, two of these, the word space and LDA models, have never been applied to verb classification before.

[3] In contrast, a top-level class like 'Transfer of Possession (Obtaining)', not only covers 25% of the gold standard, it also comprises the semantically very diverse subclasses 'Transfer of Possession (Giving)', 'Manner of Motion', and 'Emotion'.

| $N$ | Coverage of test set |
|---|---|
| 100 | 12.08% |
| 200 | 17.18% |
| 500 | 26.11% |
| 1,000 | 32.70% |
| 5,000 | 45.31% |
| 10,000 | 49.09% |
| 50,000 | 55.69% |
| 100,000 | 57.67% |

Table 1: Fraction of verb instances in the test set parameterised by LP as a function of the number of nouns $N$ included in the LP model.

### 3.1 Lexical preferences

The LP model is the simplest in our study after the baseline condition; it simply maps a noun to its own lemma. We include as a parameter of the LP model a maximum number of nouns $N$ to admit as LP tags. In this way, the LP model parameterises SCFs using only the $N$ most frequent nouns in SdeWaC; nouns beyond rank $N$ are treated as if they were unseen. Table 1 indicates what fraction of the 3 million verb instances receive SCF tags specifying one or more LPs as a function of this parameter. Note that the coverage approaches an asymptote of around 60%. This is due to the fact that noun arguments are not observed for every verb instance; many verbs' arguments are pronominal or verbal and are not treated by our SP models. Setting $N$ allows a simple way of tuning the LP model: With increasing $N$, the LP model should capture more data about verb instances, but after a point this benefit should be cancelled out by the increasing sparsity in the verb vectors.

### 3.2 Sun and Korhonen model

The SP model described in this section (SUN) was first used by Sun and Korhonen (2009) to deliver state-of-the-art verb classification performance for English; more recently, the technique was applied to successfully identify metaphor in free text (Shutova et al., 2010; Shutova et al., 2013). It uses co-occurrence counts that describe which nouns are found with which verbs in which grammatical relations; this information is used to sort the nouns into classes in a procedure almost identical to our verb clustering method described in Section 2.1.

We extract all verb instances in SdeWaC which are analysed by the SCF tagger, and count all (verb, grammatical relation, nominal argument head) triples, where the grammatical relation is subject, direct (accusative) object, indirect (dative) object, or prepositional object[4], and is listed in the verb instance's SCF tag; we undo passivisation, remove instances of auxiliary and modal verbs, and filter out those triples seen less than 10 times in the corpus.

These observations cover 60,870 noun types and 33,748,390 tokens, co-occurring with 6,705 verb types (11,426 verb-grammatical-relation types); an example is (*sprechen, obj, Wort*) ('speak' with direct object 'word', occurring 1,585 times)[5]. We represent each noun by a vector whose 11,426 dimensions are the different verb-grammatical-relation pairs; coordinates in the vector indicate the observed corpus counts. The vectors are then normalised to sum to 1, such that each represents some particular noun's discrete probability distribution over the set of verb-grammatical-relation pairs. The distance between two noun vectors is defined to be the Jensen-Shannon divergence between their probability distributions, and we partition the set of nouns into $M$ groups using hierarchical Ward's clustering.

The SP model then maps a noun to an arbitrary label indicating which of the $M$ disjoint sets that noun is to be found in (i.e., all nouns in the first noun class map to the concept label `concept1`); we employ the parameter $M$ to model SP concept granularity. As with the LP model, we use the parameter $N$ to indicate how many nouns are included in the SUN model; we search the parameter values $N = \{300, 500, 1000, 5000, 10000\}$ and $\frac{N}{M} = \{5, 10, 15, 20, 30, 50\}$.

### 3.3 Word space model

Word space models (WSMs, (Sahlgren, 2006; Turney and Pantel, 2010)) use word co-occurrence counts to represent the distributional semantics of a word. This strategy makes possible a clustering of nouns that does not depend on verbal dependencies in the first place.

[4]We have also experimented with adding features for each noun showing nominal modification features (e.g., (*schwarz, nmod, Haar*), 'hair' modified by 'black'), but these seem to hurt performance.

[5]Triples representing prepositional object relations are distinguished by preposition (e.g., the triple (*geben*, *prep-in*, *Auftrag*), 'give' with PP headed by 'in' with argument head 'contract', an idiomatic expression meaning 'to commission' something).

Dagan et al. (1999) address the problem of data sparseness for the automatic determination of word co-occurrence probabilities, which includes selectional preferences. They introduce the idea of estimating the probability of hitherto unseen word combinations using available information on words that are closest w.r.t. distributional word similarity. Following this idea, Erk (2007) and Padó et al. (2007) describe a memory-based SP model, using a WSM similarity measure to generalise the model to unseen data.

We build a WSM of German nouns and use it to partition nouns into disjoint sets, which we then employ as with the SUN model. We compute word co-occurrence counts across the whole SdeWaC corpus, using as features the 50,000 most common words in SdeWaC, skipping the first 50 most common words (i.e., we use words 50 through 50,050), with sentences as windows. We lemmatise the corpus and remove all punctuation; no other normalisation is performed. Co-occurrence counts between a word $w_i$ and a feature $c_j$ are weighted using the t-test scheme:

$$\text{ttest}(w_i, c_j) = \frac{p(w_i, c_j) - p(w_i)p(c_j)}{\sqrt{p(w_i)p(c_j)}}$$

We use a recent technique called *context selection* (Polajnar and Clark, 2014) to improve the word space model, whereby only the $C$ most highly weighted features are kept for each word vector. We set $C$ by optimising the correlation between the word space model's cosine similarity and a data set of human semantic relatedness judgements for 65 word pairs (Gurevych and Niederlich, 2005); at $C = 380$, we obtain Spearman $\rho = 0.813$ and Pearson $r = 0.707$ (human inter-annotator agreement for this data set is given as $r = 0.810$).

After this, we build a similarity matrix between all pairs of nouns using the cosine similarity, and then partition the set of $N$ nouns into $M$ disjoint classes using spectral clustering with the MNCut algorithm (Meilă and Shi, 2001). As with the SUN model, this SP model assigns labels to nouns indicating which noun class they belong to. We search the same parameter space for $N$ and $M$ as for the SUN model.

### 3.4 GermaNet

Statistical models of SPs have often used WordNet as a convenient and well-motivated inventory of concepts (e.g., Resnik (1997), Li and Abe (1998), Clark and Weir (2002)). Typically, such models make use of probabilistic treatments to determine an appropriate concept granularity separately for each predicate; we opt here for a simple model that allows more direct control over concept granularity. We take the set of concepts relevant to describing selectional preferences to be a *target set* of synsets in GermaNet (Hamp and Feldweg, 1997), and represent the target set as the set of synsets which are at some *depth* $d$ or less in the GermaNet noun hierarchy: $\{s \mid \text{depth}(s) \leq d\}$ where $\text{depth}(s)$ counts the number of hypernym links separating $s$ from the root of the hierarchy. We model concept granularity by varying $d = 1 \ldots 6$; at $d = 1$, the target set is of size 5, and at $d = 6$, it is of size 17,125. Nouns are attributed to concepts as follows: Given a noun belonging to a synset $s$, either $s$ is in the target set, or we take $s$'s lowest hypernym in the target set. For polysemous nouns, each synset listing a sense of the noun votes for a member of the target set; the noun observation is then spread over the target set using the votes as weights.

This procedure makes our GermaNet SP model a *soft clustering* over nouns (i.e., a noun can belong to more than one SP concept); a consequence of this is that a single verb occurrence in the corpus can contribute fractional counts to multiple SCF types.

### 3.5 LDA

Latent Dirichlet allocation (Blei et al., 2003) is a generative model that discovers similarities in data using latent variables; it is frequently used for topic modelling. LDA models of SPs have been proposed by Ó Séaghdha (2010) and Ritter et al. (2010); previous to this, Rooth et al. (1999) also described a latent variable model of SPs.

We implement the LDA model of selectional preferences described by Ó Séaghdha (2010). Generatively, the model produces nominal arguments to verbs as follows: For a given (verb, grammatical relation) pair $(v, r)$, (1) Sample a noun class $z$ from a from a multinomial distribution $\Phi_{v,r}$ with a Dirichlet prior parameterised by $\alpha$; (2) Sample a noun $n$ from a multinomial distribution $\Theta_z$ with a Dirichlet prior parameterised by $\beta$. Like Ó Séaghdha, we use an asymmetric Dirichlet prior for $\Phi_{v,r}$ (i.e., $\alpha$ can differ for each noun class) and a symmetric prior for $\Theta_z$ ($\beta$ is the same for each $\Theta_z$). We estimate the LDA model using the MALLET software (McCallum, 2002) using the same (verb, grammatical

relation, argument head) co-occurrence statistics used for the SUN model. We train for 1,000 iterations using the software's default parameters, allowing the LDA hyperparameters $\alpha$ and $\beta$ to be re-estimated every 10 iterations. We build models with 50 or 100 topics as a proxy to concept granularity; models include number of nouns $N$ of $\{500, 1000, 5000, 10000, 50000, 100000\}$.

As with the GermaNet-based model, the LDA model creates a soft clustering of nouns; the ability of a noun to have degrees of membership in multiple concepts might be a good way to model polysemy. We also experiment with a hard clustering version of the LDA model; to do this, we assign each noun $n$ its most likely class label $z$ using the model's estimate for $P(z|n)$.

## 4 Results

We experimented with applying the SP models to different combinations of grammatical relations (e.g., only subject, only object, subject+object, etc.), but generally obtained better results by parameterising SCF tags for all grammatical relations. Table 2 summarises the evaluation scores and parameter settings for the best-performing SP models, applied to verb arguments in all four grammatical relations (subject, direct, indirect and prepositional object)[6]. The table also indicates the number of SCF types constructed by each SP model (i.e., the number of dimensions of the vectors representing verbs).

All the SP models we compare help with automatic verb clustering. Using McNemar's test on the contingency tables underlying the $F$-scores, all models score better than the baseline at at least the $p < 0.01$ level. LDA-hard is better than the GermaNet, LDA-soft, WSM and LP models at at least the $p < 0.05$ level; SUN is better ($p \leq 0.05$) than all models except LDA-hard. All other performance differences are not statistically significant[7].

We can also demonstrate the effectiveness of the SP models with a regression analysis on the models' coverage of the test set. By varying the number of nouns $N$ included in the SP models which use this parameter (LP, SUN, WSM, LDA), or by parameterising SCF tags with SP information only for particular combinations of grammatical relations, different numbers of the verb instances in the test data will end up with SP information in their SCF tags (this is the "coverage" statistic in Table 1); with the exception of the GermaNet model, all of the SP models we examine here show positive correlation between the number of verb instances tagged for SP information and verb clustering performance. This effect is independent of parameter settings, indicating the performance benefit conferred by the SP models is robust.

### 4.1 Comparison of SP models

The GermaNet model is the least successful in our study. It achieves its best performance with a depth of 5; after this, verb clustering performance drops off again. Verb clustering using the GermaNet SP model is only slightly better than the baseline condition.

Against our expectations, the hard clustering LDA models perform better than the soft clustering ones, achieving the second highest score in our evaluation; also, in contrast to the other SP models studied in this paper, LDA performs best with fewer, coarser-grained topics. We observe that the soft clustering models produce verb vectors more than an order of magnitude longer than the hard clustering models, and suggest that simple soft clustering may be causing problems with data sparsity that interfere with verb clustering. We have also observed that the topics found by LDA do not represent polysemy as we had hoped. While some of the topics discovered by the LDA models can be easily assigned labels (e.g., body parts, people, quantities, emotions, places, buildings, tools, etc.), others are less cohesive. We found that frequent words (e.g., time, person) are generated with high probability by multiple topics in ways that do not appear to reflect multiple word senses, and that the 100-topic models exhibit this property to a greater extent. For instance, *Zeit* 'time' is highly predictive of three topics in the 50-topic models, of which only the highest-weighted topic groups time expressions together; in the 100-topic models, *Zeit* is found in six topics. Again, of these six, only the topic with the highest $\alpha$ consists of time expressions. In the 50-topic models, we find 11 topics that we cannot assign a coherent label; in the 100-topic models, there are 38 of these mismatched topics. In our work to date, we have not found that LDA models with greater numbers of topics find more

[6] Due to space constraints, we do not present here a detailed per-model study of performance as a function of parameter settings; we feel a summary to be adequate, since the relative performances of the models reflect trends across a range of parameter settings.

[7] Using a significance criterion of $p < 0.05$.

| SP model | Parameters | Granularity | $F$-score | Number of SCF types |
|---|---|---|---|---|
| SUN | 10,000 nouns | 1,000 noun classes | 39.76 | 248,665 |
| LDA (hard) | 10,000 nouns | 50 topics | 39.10 | 78,409 |
| LP | 5,000 nouns | | 38.02 | 388,691 |
| WSM | 10,000 nouns | 500 noun classes | 36.95 | 149,797 |
| LDA (soft) | 10,000 nouns | 50 topics | 35.91 | 1,524,338 |
| GermaNet | depth = 5 | 8,196 synsets | 34.41 | 851,265 |
| Baseline | | | 33.47 | 673 |

Table 2: Evaluation of the best SP models.

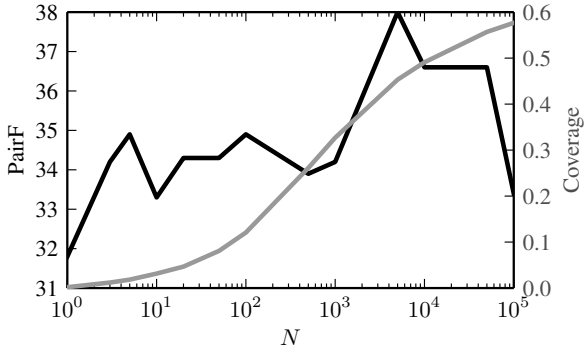


Figure 1: Verb clustering performance (black) and test set coverage (grey) of the LP model as a function of the number of nouns $N$ included in the model.

specific concepts; it is possible that this problem might be alleviated by careful filtering of the (verb, grammatical relation, noun) triples, but we leave this question to future research.

The LP model is very effective, which is surprising given its simplicity. As expected, with increasing $N$, we do observe sparsity effects which hurt verb clustering performance (see Figure 1).

Our best performing model is SUN. Our best result is obtained with 10,000 nouns (the maximum value of $N$ that we tried) in 1,000 classes, giving relatively fine-grained classes (on average 10 nouns per class). Table 3 shows some example noun classes learned by the SUN model. These include: groups with synonyms or near synonyms, often including alternate spellings of the same word (such as in the *truck* grouping); and groups of closely-related co-hyponyms, such as the body part grouping and the clothing grouping. In the latter, *bill*, *joint responsibility*, *complicity* and *inscription* are also included as things which can be *borne*, this is due to the fact that the SUN noun clustering is based on triples of verbs, grammatical relations, and nouns.

| |
|---|
| *LKW* (truck), *Lkw* (truck), *Lastwagen* (truck), *Castor* (container for highly radioactive material), *Laster* (truck), *Krankenwagen* (ambulance), *Transporter* (van), *Traktor* (tractor) |
| *Hand* (hand), *Kopf* (head), *Fuß* (foot), *Haar* (hair), *Bein* (leg), *Arm* (arm), *Zahn* (tooth), *Fell* (fur) |
| *Leiche* (corpse), *Leichnam* (body), *Schädel* (skull), *Skelett* (skeleton), *Wrack* (wreck), *Mumie* (mummy), *Trümmer* (debris) |
| *Sauna* (sauna), *Badezimmer* (bathroom), *Schwimmbad* (swimming pool), *Nachbildung* (replica), *Kamin* (fireplace), *Aufenthaltsraum* (common room), *Mensa* (cafeteria) |
| *Rechnung* (bill), *Kopftuch* (headscarf), *Uniform* (uniform), *Anzug* (suit), *Helm* (helmet), *Gewand* (garment), *Handschuh* (glove), *Mitverantwortung* (joint responsibility), *Bart* (beard), *Rüstung* (armour), *Mitschuld* (complicity), *Socke* (sock), *Jeans* (jeans), *Sonnenbrille* (sunglasses), *Aufschrift* (inscription), *Pullover* (sweater), *Weste* (vest), *Handschellen* (handcuffs), *Hörner* (horns), *Kennzeichen* (marking), *Tracht* (traditional costume), *Korsett* (corset), *Schuhwerk* (footwear), *Kopfbedeckung* (headgear), *Pelz* (fur), *Maulkorb* (muzzle) |
| *Missionar* (missionary), *Weihnachtsmann* (Santa Claus), *Selbstmordattentäter* (suicide bomber), *Bote* (messenger), *Nikolaus* (Nicholas), *Killer* (killer), *Bomber* (bomber), *Osterhase* (Easter bunny) |

Table 3: Example noun clusters in the SUN SP model.

Furthermore, there are thematically related groups (*corpse*, *body*, etc., and *sauna*, *bathroom*, etc.). All months are placed together in one 12-word group.

Some classes can be easily subdivided into separate groups, and sometimes the source for this can be guessed: For example, sports (*football*, *golf*, *tennis*) are lumped together with musical instruments (*guitar*, *piano*, *violin*) and film roles (*starring role*, *supporting role*), these all being things that can be *played*. Many groups of personal roles (such as various kinds of government ministers) are distinguished, as are diseases and medications; other groupings contain proper names or geographical locations, sometimes of surprising specificity (e.g., authors, Biblical names, philosophers, NGOs, Eastern European countries, foreign currencies, German male first names, newspapers, television channels). The last group in Table 3 shows a grouping which appears to combine two of these semantically narrow categories, in which Santa Claus and the Easter bunny are united with killers and suicide bombers.

### 4.2 Noun classes as SP concepts

The WSM SP model is not as successful as SUN, but, due to the methodological similarity between these two (SP concepts modelled as hard partitions of nouns), it affords us an opportunity to investigate the question of what properties might make for an effective noun partition.

The WSM model partitions nouns based on paradigmatic information (which sentence contexts a noun appears in), rather than SUN's use of syntagmatic information (which grammatical contexts a noun appears in). Therefore, it is perhaps not surprising that the noun classes derived by the WSM are organised thematically, and the synonym/co-hyponym structure observed in the SUN noun classes is in many cases absent (e.g., {*Pferd* (horse), *Reiter* (rider), *Stall* (stable), *Sattel* (saddle), *Stute* (mare)}; these classes can easily conflate semantic roles (e.g., Agent for rider and Location for stable), which is presumably unhelpful for representing selectional preferences.

The distribution of noun classes also differs between SUN and WSM. The largest noun class in the WSM model contains 1,076 high-frequency nouns which are semantically unrelated (*day, question, case, part, reason, kind, form, week, person, month, ...*). We suppose that these nouns are thematically "neutral" and are classed together by virtue of their usage in a wide variety of sentences. This one noun class by itself subsumes 13.6% of all noun tokens in SdeWaC. WSM also includes 56 singleton noun classes; the variance in noun class size is 2800. For comparison, in SUN, the largest noun class has 73 words, and the smallest, 2 (there are 12 of these two-word classes); noun class size variance is 37. The 73-word class in SUN does indeed appear to be a grab bag (including *gas, taboo, pioneer, mustard, spy, mafia*, and *skinhead*), but these are uncommon words and account for only 0.1% of noun tokens in SdeWaC. The next two most common classes (with some 40 nouns each) are lists of names (politicians' surnames, and male first names). The noun class in the SUN model containing the largest number of high-frequency nouns (28 nouns: *human, child, woman, man, people, Mr., mother, father, ...*) only covers 3.6% of noun usages in SdeWaC and is both semantically cohesive and intuitively useful as a SP concept.

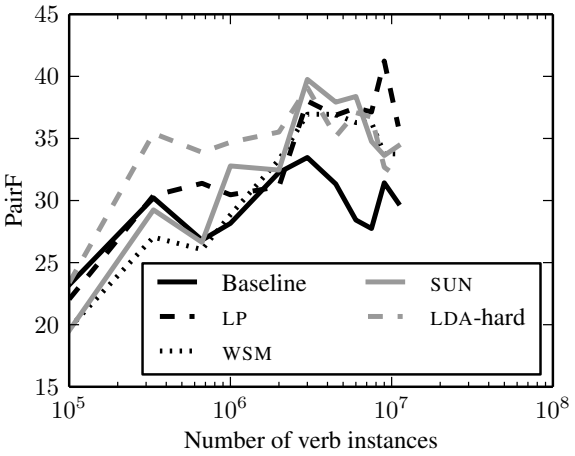


Figure 2: Verb clustering performance of SP models as a function of number of verb instances.

These issues raise the question of why the WSM model is effective at all for verb classification. We think that the larger less-related noun classes neither help nor hurt verb clustering, and we find that some of the thematic classes represent abstractions that should be useful for describing SPs. Examples include lists of body parts, countries (separate classes for Europe, Africa, Asia, etc.), diseases, human names, articles of clothing, and the group {*fruit, apple, banana, pear, strawberry*}.

### 4.3 Effects of test set size

We were curious if the success of the LP model might be due to the size of the test set preventing

sparsity from becoming a problem. To pursue this question, we take the four best performing SP models and run the verb clustering evaluation with the number of verb instances in the test set varying between 10,000 and the full SdeWaC corpus (11 million). The results are displayed in Figure 2. This graph indicates that below $3 \times 10^5$ verb instances, sparsity seems to become a problem for all models on this task, and the baseline delivers the best performance. Above this threshold, it seems that sparsity is not a major issue: LP performs fairly consistently, and is competitive with the SUN model. We attribute this to our use of the Jensen-Shannon divergence as a verb dissimilarity measure, which seems relatively robust to data sparsity. The LDA-hard model with its fewer topics seems to do quite well with fewer data; as the test set size increases, it drops off in the rankings. At the maximum number of verb instances, the best-performing models are SUN, WSM and the lexical preferences. The figure also shows that our evaluation metric is not smooth (note, e.g., the fluctuations in the baseline score). We believe that this reflects a degree of instability in the Ward's hierarchical clustering algorithm; this clustering method is greedy, and clustering errors can be expected to propagate, which might explain the jaggedness of the plot.

### 4.4 Conclusions

To conclude, we summarise the results of our analysis, using the questions formulated in the Introduction as guidelines.

First, we wanted to compare the efficiency of different classes of nouns as descriptors of selectional preference concepts. Our findings suggest that noun classes are most effective when they are semantically highly consistent, representing groups of strongly related nouns. It seems reasonable that SP concepts representing collections of synonyms would be useful for generalising observations, and should represent arguments better than simple LP. A classification of proper names (e.g., as human, corporation, country, medication) is also useful. This implies that we can expect features such as ANIMATE to be shared by all members of a noun cluster.

Second, we were interested in the appropriate granularity of selectional preference concepts. In our evaluation, we have observed a tendency for smaller, more specific noun classes to be superior; this holds because data sparsity is not a problem in our experiment. Beyond this finding, we would have liked to present a direct juxtaposition of different models on "granularity" but this is difficult: We have not yet identified a strong abstraction of granularity from the proxies we use (e.g., GermaNet depth, or SUN's $N/M$).

Finally, which methods of classifying nouns into concepts are most effective at capturing selectional preferences for verb clustering? In our experiments, the SUN and LDA-hard models proved to be more effective than lexical preferences, supporting our primary hypothesis that some level of SP concept granularity above the lexical level is desirable for verb clustering. On the other hand, the LP model is only slightly worse than SUN and LDA-hard, making it attractive because it is so simple. As we have shown, the potential data sparsity issues with LP can be alleviated by judiciously choosing the value of the $N$ parameter that controls the number of nouns included in the model. In addition, comparing the SUN and WSM models, and observing the performance of the LDA-hard method, we conclude that inducing noun classes using syntagmatic information is more effective than using paradigmatic relations.

## 5 Related work

In this study, we have looked at the utility of selectional preferences for automatic verb classification. Some previous research has followed this line of inquiry, though prior studies have not compared alternative methods of modelling SPs. Schulte im Walde (2006) presented a detailed examination of parameters for $k$-means-based verb clustering in German, using the same gold standard that we employ here. She reports on the effects of adding SP information to a SCF-based verb clustering using 15 high-level GermaNet synsets as SP concepts; SP information for some combinations of grammatical relations improves clustering performance slightly, but neither are the effects consistent, nor is the improvement delivered by the SP model over the SCF-based baseline statistically significant. Schulte im Walde et al. (2008) used expectation maximisation to induce latent verb clusters from the British National Corpus while simultaneously building a tree cut model of SPs on the WordNet hierarchy using a minimum description length method; their evaluation focuses on the induced soft verb clusters, reporting the model's estimated perplexity of (verb, grammatical relation, argument head) triples. The

SPs are described qualitatively by presenting two example cases. Sun and Korhonen (2009) study the effect of adding selectional preferences to a subcategorisation-based verb clustering in English using the SUN model (see Section 3.2). They demonstrate that adding SPs to the SCF preference data leads to the best results on their two clustering evaluations; overall, their best results come from using SP information only for the subject grammatical relation. They employ coarse SP concepts (20 or 30 noun clusters) which capture general semantic categories (*Human, Building, Idea*, etc.).

Selectional preferences are usually evaluated either from a word sense disambiguation standpoint using pseudo-words (Chambers and Jurafsky, 2010), or in terms of how acceptable an argument is with a verb, via regression against human plausibility judgements. Several studies have compared SP methodologies from the latter perspective. These include Brockmann and Lapata (2003), who compared three GermaNet-based models of SP, showing that different models were most effective for describing different grammatical relations; Ó Séaghdha (2010), who compared different LDA-based models of SP, showing these to be effective for a variety of grammatical relations; and Ó Séaghdha and Korhonen (2012), who show that WordNet tree cut models, LDA, and a hybrid LDA-WordNet model are effective for describing verb-object relations.

## 6 Future work

Our GermaNet model delivered disappointing performance in this study; we would be interested in seeing whether a more sophisticated implementation such as the tree cut model of Li and Abe (1998) would be more competitive. We also would like to explore alternative noun clustering methods such as CBC (Pantel and Lin, 2002) and Brown clusters (Brown et al., 1992), which were not covered in this work; these would fit easily into our SP evaluation paradigm. More challenging would be a verb classification-based evaluation of the SP models of (Rooth et al., 1999) and (Schulte im Walde et al., 2008), which use expectation maximisation to simultaneously cluster verbs into verb classes and nominal arguments into noun classes; these approaches are not compatible with the evaluation framework we have used here. Finally, the SP model of Bergsma et al. (2008) has also achieved impressive results on a number of tasks, but has not been investigated for use in verb classification.

Our verb clustering evaluation in this work has matched $K$, the number of clusters found by Ward's method, to the number of classes in the gold standard. Since the number of clusters has an influence on the quality of the ensuing semantic classification (Schulte im Walde, 2006, page 180f.), we will also be running our experiments with different settings of $K$ to explore whether this also influences the overall results of our evaluation.

## References


Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 59–68.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Carsten Brockmann and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, pages 27–34.

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

Nathanael Chambers and Daniel Jurafsky. 2010. Improving the use of pseudo-words for evaluating selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 445–453.

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

Ido Dagan, Lillian Lee, and Fernando C.N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 216–223.

Eva Esteve Ferrer. 2004. Towards a semantic classification of Spanish verbs based on subcategorisation information. In *Proceedings of the Student Research Workshop at the Annual Meeting of the Association for Computational Linguistics*, pages 37–42.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC - A corpus of parsable sentences from the Web. In *Language processing and knowledge in the Web*, pages 61–68. Springer, Berlin, Heidelberg.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 273–283.

Iryna Gurevych and Hendrik Niederlich. 2005. Computing semantic relatedness in German with revised information content metrics. In *Proceedings of "OntoLex 2005 - Ontologies and Lexical Resources" IJCNLP'05 Workshop*, pages 28–33.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet: A lexical-semantic net for German. In *Proceedings of ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, pages 172–182.

Jerrold J. Katz and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39:170–210.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2008. The choice of features for classification of verbs in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 449–456.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, Chicago.

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

Jianguo Li and Chris Brew. 2008. Which are the best features for automatic verb classification. In *Proceedings of ACL-08: HLT*, pages 434–442.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Andrew McCallum. 2002. MALLET: A machine learning for language toolkit.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Marina Meilă and Jianbo Shi. 2001. A random walks view of spectral segmentation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Diarmuid Ó Séaghdha and Anna Korhonen. 2012. Modelling selectional preferences in a lexical hierarchy. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, pages 170–179.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444.

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 400–409.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619.

Tamara Polajnar and Stephen Clark. 2014. Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238.

Philip Resnik. 1993. *Selection and information: A class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania.

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.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434.

Will Roberts, Markus Egg, and Valia Kordoni. 2014. Subcategorisation acquisition from raw text for a free word-order language. In *Proceedings of the*

*14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 298–307.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 104–111.

Magnus Sahlgren. 2006. *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm University.

Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 496–504.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 747–753.

Sabine Schulte im Walde. 2002. A subcategorisation lexicon for German verbs induced from a lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC)*, pages 1351–1357.

Sabine Schulte im Walde. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Sabine Schulte im Walde. 2009. The induction of verb frames and verb classes from corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics: An international handbook*, volume 2, chapter 44, pages 952–971. Mouton de Gruyter, Berlin.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Mark Stevenson and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 638–647.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Verb class discovery from rich syntactic data. In *Proceedings of the Ninth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 16–27.

Lin Sun. 2012. *Automatic induction of verb classes using clustering*. Ph.D. thesis, University of Cambridge, Cambridge.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Joe H. Ward, Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Yorick Wilks. 1975. An intelligent analyzer and understander of English. *Communications of the ACM*, 18(5):264–274.

Yorick Wilks. 1978. Making preferences more active. *Artificial Intelligence*, 11(3):197–223.