Subcategorisation Acquisition from Raw Text for a Free Word-Order Language

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

Abstract

We describe a state-of-the-art automatic system that can acquire subcategorisation frames from raw text for a free word-order language. We use it to construct a subcategorisation lexicon of German verbs from a large Web page corpus. With an automatic verb classification paradigm we evaluate our subcategorisation lexicon against a previous classification of German verbs; the lexicon produced by our system performs better than the best previous results.

1 Introduction

We introduce a state-of-the-art system for the acquisition of subcategorisation frames (SCFs) from large corpora, which can deal with languages with very free word order. The concrete language we treat is German; its word order variability is illustrated in (1)–(4), all of which express the sentence *The man gave the old dog a chop*:

(1) Dem alten Hund gab der Mann ein Schnitzel.

(2) Ein Schnitzel gab dem alten Hund der Mann.

(3) Ein Schnitzel gab der Mann dem alten Hund.

(4) Der Mann gab dem alten Hund ein Schnitzel.

On the basis of raw text, the system can be used to build extensive SCF lexicons for German verbs. Subcategorisation means that lexical items require specific obligatory concomitants or arguments; we focus on verb subcategorisation. E.g., the verb *geben* 'give' requires three arguments, the nominative subject *der Mann* 'the man', the dative indirect object *dem alten Hund* 'the old dog', and the accusative direct object *ein Schnitzel* 'a chop'. Other syntactic items may be subcategorised for, too, e.g. both *stellen* and its English translation *put* subcategorise for subject, direct object, and a prepositional phrase (PP) like *on the shelf*:

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

Subcategorisation frames describe a combination of arguments required by a specific verb. The set of SCFs for a verb is called its *subcategorisation preference*. Our system follows much previous work by counting PPs that accompany the verb among its complements, even though they are not obligatory (so-called 'adjuncts'), because PP adjuncts are excellent clues to a verb's semantics (Sun et al., 2008). However, nominal and clausal adjuncts do not count as verbal complements.

SCF information can benefit all applications that need information on predicate-argument structure, e.g., parsing, verb clustering, semantic role labelling, or machine translation. Automatic acquisition of SCF information with minimal supervision is also crucial to construct useful resources quickly.

The main innovation of the presented new system is to address two challenges simultaneously, viz., SCF acquisition from *raw text* and the focus on languages with a *very free word order*. With this system, we create an SCF lexicon for German verbs and evaluate this lexicon against a previously published manual verb classification, showing better performance than has been reported until now.

After an overview of previous work on SCF acquisition in Section 2, Section 3 describes our subcategorisation acquisition system, and Section 4 the SCF lexicon that we build using it. In Sections 5 and 6 we evaluate the SCF lexicon on a verb classification task and discuss our results; Section 7 then concludes with directions for future work.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 298–307,
Gothenburg, Sweden, April 26-30 2014.

2 Previous work

To date, research on SCF acquisition from corpora has mostly targeted English. Brent and Berwick (1991) detect five SCFs by looking for attested contexts where argument slots are filled by closed-class lexical items (pronouns or proper names). Briscoe and Carroll (1997) detect 163 SCFs with a system that builds an SCF lexicon whose entries include the relative frequency of SCF classes. Potential SCF patterns are extracted from a corpus parsed with a dependency-based parser, and then filtered by hypothesis testing on binomial frequency data. Korhonen (2002) refines Briscoe and Carroll (1997)'s system using back-off estimates on the WordNet semantic class of the verb's predominant sense, assuming that semantically similar verbs have similar SCFs, following Levin (1993). Some current statistical methods for Semantic Role Labelling build models that also capture subcategorisation information, e.g., Grenager and Manning (2006). Schulte im Walde (2009) offers a recent survey of the SCF acquisition literature.

SCF acquisition is also an important step in the automatic semantic role labelling (Grenager and Manning, 2006; Lang and Lapata, 2010; Titov and Klementiev, 2012). Semantic roles of a verb describe the kind of involvement of entities in the event introduced by the verb, e.g., as agent (active, often not affected by the event) or patient (passive, often affected). On the basis of these SCFs, semantic roles can be assigned due to the interdependence between semantic roles and their syntactic realisations, called *Argument Linking* (Levin, 1993; Levin and Rappaport Hovav, 2005).

Acquiring SCFs for languages with a very fixed word order like English needs only a simple syntactic analysis, which mainly relies on the predetermined sequencing of arguments in the sentence, e.g., Grenager and Manning (2006). When word order is freer, the analysis gets more complicated, and must include a full syntactic parse.

What is more, German is a counterexample to Manning's (1993) expectation that freedom of word order should be matched by an increase in case and/or agreement marking. This is due to a very high degree of syncretism (identity of word forms) in German paradigms for nouns, adjectives, and determiners. E.g., the noun *Auto* 'car' has only two forms, *Auto* for nominative, dative, and accusative singular, and *Autos* for genitive singular and all four plural forms. This is in contrast to some other free word order languages for which SCF acquisition has been studied, like Modern Greek (Maragoudakis et al., 2000) and Czech (Sarkar and Zeman, 2000). A one-many relation between word forms and case is also one of the problems for SCF acquisition in Urdu (Ghulam, 2011).

For German, initial studies used semi-automatic techniques and manual evaluation (Eckle-Kohler, 1999; Wauschkuhn, 1999). The first automatic subcategorisation acquisition system for German is described by Schulte im Walde (2002a), who defined an SCF inventory and manually wrote a grammar to analyse verb constructions according to these frames. A lexicalised PCFG parser using this grammar was trained on 18.7 million words of German newspaper text; the trained parser model contained explicit subcategorisation frequencies, which could then be extracted to construct a subcategorisation lexicon for 14,229 German verbs. This work was evaluated against a German dictionary, the *Duden Stilwörterbuch* (Schulte im Walde, 2002b).

Schulte im Walde and Brew (2002) used the subcategorisation lexicon created by the system to automatically induce a set of semantic verb classes with an unsupervised clustering algorithm. This clustering was evaluated against a small manually created semantic verb classification. Schulte im Walde (2006) continues this work using a larger manual verb classification. The SCFs used in this study are defined at three levels of granularity. The first level (38 different SCFs) lists only the complements in the frame; the second one adds head and case information for PP complements (183 SCFs). The third level examined the effect of adding selectional preferences, but results were inconclusive.

A recent paper (Scheible et al., 2013) describes a system similar to ours, built on a statistical dependency parser, and using some of the same kinds of rules as we describe in Section 3.1; this system is evaluated in a task-based way (e.g., to improve the performance of a SMT system) and cannot be directly compared to our system in this paper.

3 The SCF acquisition system

This section describes the first contribution of this paper, a state-of-the-art subcategorisation acquisition system for German. Its core component is a rule-based SCF tagger which operates on phrase structure analyses, as delivered by a statistical parser. Given a parse of a sentence, the tagger assigns each finite verb in the sentence an SCF type.

We use the SCF inventory of Schulte im Walde (2002a), which includes complements like `n` for nominative subject, `a` for accusative direct object, `d` for dative indirect object, `r` for reflexive pronoun, and `x` for expletive *es* ('it') subject. Clausal complements can be infinite (`i`); finite ones can have the verb in second position (`S-2`) or include the complementiser *dass* 'that' (`S-dass`). Complements can be combined as in `na` (transitive verb); for PPs in SCFs, the head is specified, e.g., `p:für` for PP complements headed by *für* 'for'[1].

Due to the free word order, simple phrase structure like that used for analysis of English is not enough to specify the syntax of German sentences. Therefore we use the annotation scheme in the manually constructed German treebanks NEGRA and TIGER (Skut et al., 1997; Brants et al., 2002), which decorate parse trees with edge labels specifying the syntactic roles of constituents. We automatically annotate the parse trees from our statistical parser using a simple machine learning model.

In the next section, we illustrate the operation of the SCF tagger with reference to examples; then in Section 3.2 we describe our edge labeller.

3.1 The SCF tagger

The SCF tagger begins by collecting complements co-occurring with a verb instance using the phrase structure of the sentence. In our system, we obtain phrase structure information for unannotated text using the Berkeley Parser (Petrov et al., 2006), a statistical unlexicalised parser trained on TIGER. Fig. 1 illustrates the phrase structure analysis and edge labels in the TIGER corpus for (6):

(6) *Das hielte ich für moralisch außerordentlich fragwürdig.*
'I'd consider that morally extremely questionable'.

Its finite verb *hielte* (from *halten* 'hold') has three complements, the subject *ich* 'I', edge-labelled with `SB`, the direct object *das* 'that', labelled with `OA`, and a PP headed by *für* 'for' (`MO` stands for 'modifier'). After collecting complements, the SCF tagger uses this edge label information to determine the complements' syntactic roles, and assigns the verb the corresponding SCF; in the case of *halten* above, the SCF is `nap:für`.

The rule-based SCF tagger handles auxiliary and modal verb constructions, passive alternations, separable verb prefixes, and raising and control constructions. E.g., the subject *sie* 'they' of *anfangen* 'begin' in (7) doubles as the subject of its infinite clausal complement; hence, it shows up in the SCF of the complement's head *geben* 'give', too:

(7) *Sie fingen an, mir Stromschläge zu geben.*
'They started to give me electric shocks.'

The tagger also handles involved cases with many complements, including PPs and clauses as in (8). As the SCF inventory allows at most three complements in an SCF, such cases call for prioritising of verbal complements (e.g., subjects, objects, and clausal complements are preferred over PP complements). Consequently, the main verb *empfehlen* 'recommend' in (8), which has a subject, a dative object, a PP, and an infinitival clausal complement, is assigned the SCF `ndi`. Another challenging task which relies on edge label information is filtering out clausal adjuncts (relative clauses and parentheticals) so as not to include them in SCFs.

(8) *[$_{PP}$ Am Freitag] empfahl [$_{NP:Nom}$ der Aufsichtsrat] [$_{NP:Dat}$ den Aktionären], [$_{S}$ das Angebot abzulehnen].*
'On Friday the board of directors advised shareholders to turn down the offer.'

The 17 rules of the SCF tagger are simple; most of them categorise the complements of a specific verb instance; e.g., if a nominal complement to the verb is edge-labelled as a nominative subject, add `n` to the verb's SCF, unless the verb is in the passive, in which case add `a` to the SCF.

Our system was optimised by progressively refining the SCF tagger's rules through manual error analysis on sentences from TIGER. The result is an automatic SCF tagger that is resilient to variations in sentence structure and is firmly based on linguistically motivated knowledge. As a test case for its linguistic soundness, we chose the perfect parses in the TIGER treebank and found that the tagger is very accurate in capturing subcategorisation information inherent in these data.

3.2 The edge labeller

To obtain edge label information for the parses delivered by the Berkeley Parser, we built a novel machine learning classifier to annotate parse trees

[1] We digress from Schulte im Walde's original SCF inventory in that we do not indicate case information in PPs.

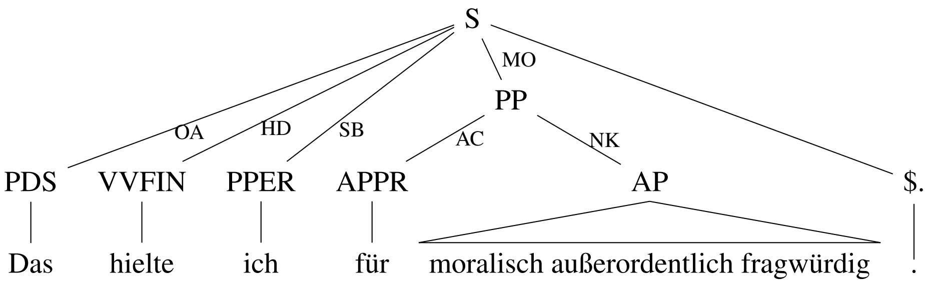

Figure 1: Edge labels in the TIGER corpus.

with TIGER edge label information. This edge labeller is a maximum entropy (multiclass logistic regression) model built using the Stanford Classifier package[2]. We include features such as:

- The part of speech of the complement;
- The first word of the complement;
- The lexical head of the complement;
- N-grams on the end of the lexical head of the complement;
- The kind of article of a complement;
- The presence or absence of specific article forms in other complements to the same verb;
- Position of the complement with respect to a reflexive pronoun in the sentence;
- The lemmatised form of the verb governing the complement (i.e., the verb on which the complement depends syntactically);
- The clause type of the governing verb; and,
- Active or passive voice of the governing verb.

We do no tuning and use the software's default hyperparameters (L2 regularisation with $\sigma = 3$).

This classifier was trained from edge label data extracted from the NEGRA and TIGER corpora; our training set contained 300,000 samples (approximately 25% from NEGRA and 75% from TIGER). On a held-out test set of 10% (containing 34,000 samples), the classifier achieves a final F-score of 95.5% on the edge labelling task.

The edge labeller makes the simplifying assumption that verbal complements can be labelled independently. Consequently, it tends to annotate multiple complements as subject for each verb. This has to do with the numerical dominance of subjects, which make up about 40% of all verb complements, more than three times the number of the next most common complement type (direct object).

Therefore we first collect all possible labels with associated probabilities that the edge labeller assigns to each complement of a verb. We then choose the set of labels with the highest probability that includes at most one subject and at most one accusative direct object for the verb, assuming that the joint probability of a set of labels is the product of the individual label probabilities.

We use our edge labeller in this work for morphological disambiguation of nominals and for identifying clausal adjuncts, but the edge labeller is a standalone reusable component, which might be equally well be used to mark up parse trees for, e.g., a semantic role labelling system.

4 The subcategorisation lexicon

With the system described in Sec. 3, we build a German subcategorisation lexicon that collects counts of ⟨lemma, SCF⟩ on deWaC (Baroni et al., 2009), a corpus of text extracted from Web search results, with 10^9 words automatically POS-tagged and lemmatised by the TreeTagger (Schmid, 1994). A subset of this corpus, SdeWaC (Faaß and Eckart, 2013), has been preprocessed to include only sentences which are maximally parsable; this smaller corpus includes 880 million words in 45 million sentences. We parsed 3 million sentences (80 million words) of SdeWaC; after filtering out those verb lemmas seen only five times or fewer in the corpus, we are left with statistics on 8 million verb instances, representing 9,825 verb lemmas.

As a concrete example for the resulting SCF lexicon, consider the entry for *sprechen* 'talk' in Fig. 2, which occurs 16,254 times in our SCF lexicon.

Sprechen refers to a conversation with speaker, hearer, topic, message, and code: Speakers are expressed by nominative NPs, hearers, by *mit*-, *bei*- or *zu*-PPs, topics, by *von*- and *über*-PPs. The code is expressed in *in*-PPs, and the message, by accusative NPs (*einige Worte sprechen* 'to say a few words'), main-clause complements or subordinate *dass* ('that') sentences. Other uses of the verb are

[2] http://nlp.stanford.edu/software/classifier.shtml

`np:von` (2715), `n` (2696), `na` (1380), `np:mit` (1247), `np:in` (1132), `nS-2` (1064), `np:über` (853), `np:für` (695), `nS-dass` (491), `np:zu` (307), `nap:in` (280), `nap:von` (275), `ni` (261), `np:bei` (212), `np:gegen` (192), `np:an` (186), `naS-2` (172), `np:aus` (168), `np:auf` (112), `nap:über` (112)

Figure 2: SCF lexicon for *sprechen*

figurative , e.g., *sprechen gegen* 'be a counterargument to'. As the distinction between arguments and adjuncts is gradual in our system, some adjunct patterns appear in the lexicon, too, but only with low frequency, e.g., `np:auf`, in which the *auf*-PP expresses the setting of the conversation, as in *auf der Tagung sprechen* 'speak at the convention'.

For reference, we also constructed an SCF lexicon from the NEGRA and TIGER corpora, which together comprise about 1.2 million words. This SCF lexicon contains statistics on 133,897 verb instances (5,316 verb lemmas). While the manual annotations in NEGRA and TIGER mean that this SCF lexicon has virtually no noise, the small size of the corpora results in problems with data sparsity and negatively impacts the utility of this resource (see discussion in Section 6.2).

5 Automatic verb classification

The remainder of the paper sets out to establish the relevance of our SCF acquisition system by comparison to previous work. As stated in Sec. 2, the only prior automatic German SCF acquisition system is that of Schulte im Walde (2002a), which was evaluated directly against an electronic version of a large dictionary; as this is not an open access resource, we cannot perform a similar evaluation.

We opt therefore to use a task-based evaluation to compare our system directly with Schulte im Walde's, and leave manual evaluation for future work. We refer back to the experiment set up by Schulte im Walde (2006) to automatically induce classifications of German verbs by clustering them on the basis of their SCF preferences as listed in her SCF lexicon. By casting this experiment as a fixed task, we can compare our system directly to hers. The link between subcategorisation and verb semantics is linguistically sound, due to the interdependence between verb meanings and the number and kinds of their syntactic arguments (Levin, 1993; Levin and Rappaport Hovav, 2005). E.g., only transitive verbs that denote a change of state like *cut* and *break* enter in the middle construction (*The bread cuts easily.*), with the patient or theme argument appearing as the syntactic subject. Thus, verbs whose SCF preferences show such an alternation can be predicted to denote a change of state.

We adopt the automatic verb classification paradigm to evaluate our system, replicating Schulte im Walde's (2006) experiment to the best of our ability. We argue that by evaluating our SdeWaC SCF lexicon described in the previous section, we simultaneously evaluate our subcategorisation acquisition system; this technique also allows us to demonstrate the semantic relevance of our SCF lexicon. Section 5.1 introduces the manual verb classification we use as a gold standard and Section 5.2 describes our unsupervised clustering technique. Our evaluation of the clustering against the gold standard then follows in Section 6.

5.1 Manual verb classifications

The semantic verb classification proposed by Schulte im Walde (2006, page 162ff.), hereafter SiW2006, comprises 168 high- and low-frequency verbs grouped into 43 semantic classes, with between 2 and 7 verbs per class. Examples of these classes are Aspect (e.g., *anfangen* 'begin'), Propositional Attitude (e.g., *denken* 'think'), Transfer of Possession (Obtaining) (e.g., *bekommen* 'get'), and Weather (e.g., *regnen* 'rain'). Some of the classes are subclassified[3], e.g., Manner of Motion, with the subclasses Locomotion (*klettern* 'climb'), Rotation (*rotieren* 'rotate'), Rush (*eilen* 'hurry'), Vehicle (*fliegen* 'fly'), and Flotation (*gleiten* 'glide').

These classes are related to Levin classes in that some are roughly equivalent to a Levin class (e.g., Aspect and Levin's Begin class), others are subgroups of Levin classes, e.g., Position is a subgroup of Levin's Dangle class; finally, some classes lump together Levin classes, e.g., Transfer of Possession (Obtaining) combines Levin's Get and Obtain classes. This shows that these classes could be integrated into a large-scale classification of German verbs in the style of Levin (1993).

5.2 Clustering

From the counts of $\langle \text{lemma}, \text{SCF} \rangle$ in the SCF lexicon, we can estimate the conditional probability that a particular verb v appears with an SCF f:

[3] For the purpose of our evaluation, we disregard class-subclass relations and consider subclasses as separate entities.

$P(\text{scf} = f|\text{lemma} = v)$. We smooth these conditional probability distributions by backing off to the prior probability $P(\text{scf})$ (Katz, 1987).

With these smoothed conditional probabilities, we cluster verbs with k-means clustering (Forgy, 1965), a hard clustering technique, which partitions a set of objects into k clusters. The algorithm is initialised with a starting set of k cluster centroids; it then proceeds iteratively, first assigning each object to the cluster whose centroid is closest under some distance measure, and then calculating new centroids to represent the centres of the updated clusters. The algorithm terminates when the assignment of objects to clusters no longer changes.

$$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} \quad (9)$$

$$\text{irad}(p,q) = D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2}) \quad (10)$$

$$\text{skew}(p,q) = D(p\|\alpha q + (1-\alpha)p) \quad (11)$$

In our experiments, verbs are represented by their conditional probability distributions over SCFs. As distance measures, we use two variants of the Kullback-Leibler divergence (9), a measure of the dissimilarity of two probability distributions. The KL divergence from p to q is undefined if at some point q but not p is zero, so we use measures based on KL without this problem, viz., the *information radius* (aka Jensen-Shannon divergence, a symmetric metric, (10)), as well as *skew divergence* (an asymmetric dissimilarity measure which smoothes q by interpolating it to a small degree with p, (11)), where we set the interpolation parameter to be $\alpha = 0.9$, to make our results comparable to Schulte im Walde's (2006)[4].

As mentioned, the k-means algorithm is initialised with a set of cluster centroids; in this study, we initialise the centroids by random partitions (each of the n objects is randomly assigned to one of k clusters, and the centroids are then computed as the means of these random partitions). Because the random initial centroids influence the final clustering, we repeat the clustering a number of times.

We also initialise the k-means cluster centroids using agglomerative hierarchical clustering, a deterministic iterative bottom-up process. Hierarchical clustering initially assigns verbs to singleton clusters; the two clusters which are "nearest" to each other are then joined together, and this process is repeated until the desired number of clusters is obtained. Hierarchical clustering is performed to group the verbs into k clusters; the centroids of these clusters are then used to initialise the k-means algorithm. While there exist several variants of hierarchical clustering, we use Ward's method (Ward, Jr, 1963) for merging clusters, which attempts to minimise the variance inside clusters; Ward's criterion was previously found to be the most effective hierarchical clustering technique for verb classification (Schulte im Walde, 2006).

6 Evaluation

This section presents the results of evaluating the unsupervised verb clustering based on our SCF lexica against the gold standard described in Sec. 5.1.

6.1 Results

We use two cluster purity measures, defined in Fig. 3; we intentionally target our numerical evaluations to be directly comparable with previous results in the literature. As k-means is a hard clustering algorithm, we consider a clustering $\mathcal{C}$ to be an equivalence relation that partitions n verbs into k disjoint subsets $\mathcal{C} = \{C_1, \ldots, C_k\}$.

The first of these purity measures, *adjusted Rand index* (Rand_a in Eq. (12)) judges clustering similarity using the notion of the overlap between a cluster C_i in a given clustering $\mathcal{C}$ and a cluster G_j in a gold standard clustering $\mathcal{G}$, this value being denoted by $\mathcal{CG}_{ij} = |C_i \cap G_j|$; values of Rand_a range between 0 for chance and 1 for perfect correlation. The other metric, the *pairwise F-score* (PairF, Eq. (13)), operates by constructing a contingency table on 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 in clustering a pair of verbs together or separately, this is a "correct" answer; by extension, information retrieval measures such as precision (P) and recall (R) can be computed.

Table 1 shows the performance of our SCF lexica, evaluated against the SiW2006 gold standard. The random baseline is given by PairF $= 2.08$ and $\text{Rand}_a = -0.004$ (calculated as the average of 50 random partitions). The optimal baseline is PairF $= 95.81$ and $\text{Rand}_a = 0.909$, calculated by evaluating the gold standard against itself. As the gold standard includes polysemous verbs, which belong

[4] Schulte im Walde (2006) takes $\alpha = 0.9$ although Lee (1999) recommends $\alpha = 0.99$ or higher values in her original description of skew divergence.

$$\text{Rand}_a(\mathcal{C},\mathcal{G}) = \frac{\sum_{i,j}\binom{\mathcal{CG}_{ij}}{2} - \left[\sum_i \binom{|C_i|}{2}\sum_j \binom{|G_j|}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{|C_i|}{2} + \sum_j \binom{|G_j|}{2}\right] - \left[\sum_i \binom{|C_i|}{2}\sum_j \binom{|G_j|}{2}\right]/\binom{n}{2}} \tag{12}$$

$$\text{PairF}(\mathcal{C},\mathcal{G}) = \frac{2P(\mathcal{C},\mathcal{G})R(\mathcal{C},\mathcal{G})}{P(\mathcal{C},\mathcal{G}) + R(\mathcal{C},\mathcal{G})} \tag{13}$$

Figure 3: Evaluation metrics used to compare clusterings to gold standards.

Data Set	Eval	Distance	Manual	Random Best	Random Mean	Ward
Schulte im Walde	PairF	IRad	40.23	1.34 → 16.15	13.37	17.86 → 17.49
		Skew	47.28	2.41 → 18.01	14.07	15.86 → 15.23
	Rand$_a$	IRad	0.358	0.001 → 0.118	0.093	0.145 → 0.142
		Skew	0.429	−0.002 → 0.142	0.102	0.158 → 0.158
NEGRA/TIGER	PairF	IRad	30.77	2.06 → 14.67	12.39	16.13 → 15.52
		Skew	40.19	3.47 → 12.95	11.48	14.05 → 14.31
	Rand$_a$	IRad	0.281	0.000 → 0.122	0.094	0.134 → 0.129
		Skew	0.382	−0.015 → 0.102	0.089	0.112 → 0.114
SdeWaC	PairF	IRad	42.66	1.62 → 20.36	18.26	26.94 → 27.50
		Skew	50.38	2.99 → 20.75	17.80	24.60 → 24.94
	Rand$_a$	IRad	0.387	−0.006 → 0.167	0.146	0.232 → 0.238
		Skew	0.465	0.008 → 0.170	0.143	0.208 → 0.211

Table 1: Evaluation of the NEGRA/TIGER and SdeWaC SCF lexica using the SiW2006 gold standard.

to more than one cluster, the optimal baseline is calculated by randomly picking one of their senses; the average is then taken over 50 such trials.

We cluster using $k = 43$, matching the number of clusters in the gold standard. Of the 168 verbs in SiW2006, 159 are attested in NEGRA and TIGER (17,285 instances), and 167 are found in SdeWaC (1,047,042 instances)[5].

We report the results using k-means clustering initialised under a variety of conditions. “Manual” shows the quality of the clustering achieved when initialising k-means with the gold standard classes. We also initialise clustering 10 times using random partitions. For the best clustering[6] in these 10, “Random Best” shows the evaluation of both the starting random partition and the final clustering found by k-means; “Random Mean” shows the average cluster purity of the 10 final clusterings. “Ward” shows the evaluation of the clustering initialised with centroids found by hierarchical clustering of the verbs using Ward’s method. Again, both the initial partition found by Ward’s method and the k-means solution based on it are shown.

For comparison, we list the results of Schulte im Walde (2006, p. 174, Table 7) for the second level of SCF granularity, with PP head and case information (see Sec. 2 for Schulte im Walde’s analysis). While this seems the most appropriate comparison to draw, since we also collect statistics about PPs, it is ambitious because, as noted in Section 3, our SCF lexica lack case information about PPs.[7] Compared to Schulte im Walde’s numbers, the NEGRA/TIGER SCF lexicon scores significantly worse on the PairF evaluation metric under all conditions, and also on the Rand$_a$ metric using the skew divergence measure (Rand$_a$/IRad is not significantly different). The SdeWaC SCF lexicon scores better on all metrics and conditions; these results are significant at the $p < 0.001$ level[8].

[5] Verbs missing from the clustering reduce the maximum achievable cluster purity score.

[6] Specifically, we take the clustering result with the minimum intra-cluster distance (not the clustering result with the best performance on the gold standard).

[7] PP case information is relevant for prepositions that can take both locative and directional readings, as in *in der Stadt* (dative) ‘in town’ und *in die Stadt* (accusative) ‘to town’.

[8] Statistical significance is calculated by running repeated k-means clusterings with random partition initialisation and evaluating the results using the relevant purity metrics. These repeated clustering scores represent a random variable (a func-

6.2 Discussion

Sec. 6.1 compared the SCF lexicon created using SdeWaC with the lexicon built by Schulte im Walde (2002a), showing that our lexicon achieves significantly better results on the verb clustering task. We interpret this to be indicative of a more accurate subcategorisation lexicon, and, by extension, of a more accurate SCF acquisition system.

We attribute this superior performance primarily to our use of a statistical parser as opposed to a hand-written grammar. This design choice has several advantages. First, the parser delivers robust syntactic analyses, which we can expect to be relatively domain-independent. Second, we make no prior assumptions about the variety of subcategorisation phenomena that might appear in text, decoupling the identification of SCFs from the ability to parse natural language. Third, the fact that our parser and edge labeller are trained on the 800,000 word NEGRA/TIGER corpus means that we benefit from the linguistic expertise that went into building that treebank. Our use of off-the-shelf tools (the parser and our simple yet effective machine learning model describing edge label information) makes our system considerably simpler and easier to implement than Schulte im Walde's. We see our system as more easily extensible to other languages for which there is a parser and an initial syntactically annotated corpus to train the edge labeller on.

The NEGRA/TIGER SCF lexicon performs not as well on the verb clustering evaluations, as fewer verbs are attested in NEGRA/TIGER compared to the SdeWaC SCF lexicon and gold standard clusterings. Data sparsity can be a problem in SCF acquisition; all other factors being equal, using more data to construct an SCF lexicon should make patterns in the language more readily visible and reduce the chance of missing a particular lemma-SCF combination accidentally. A secondary effect is that models of verb subcategorisation preferences like the ones used here can be more precisely estimated as the counts of observed verb instances increase, particularly for low-frequency verbs.

Error analysis of our SCF lexicon reveals low counts of expletive subjects. The edge labeller is supposed to annotate semantically empty subjects (*es*, 'it') as expletive; for clusterings examined in Sec. 5.1, this would affect weather verbs (e.g., *es regnet*, 'it's raining'). However, in our SdeWaC SCF lexicon, expletive subjects are clearly underrepresented. Our SCF lexicon built on TIGER, where expletive subjects are systematically labelled, has the SCF `xa` as the most common SCF for the verb *geben* (in *es gibt* 'there is'). In contrast, in our SdeWaC SCF lexicon, the most common SCF is the transitive `na`, with `xa` in seventh place. I.e., the edge labeller does not identify all expletive subjects, which is due to the fact that expletive subjects are syntactically indistinguishable from neuter pronominal subjects, so the edge labeller does not have a rich feature set to inform it about this category. But since, statistically, expletive pronouns make up less than 1% of subjects in TIGER, the prior probability of labelling a constituent as expletive is very low. Due to these figures, we do not expect this issue to seriously impact the quality of our verb classification evaluations.

7 Future work

In this paper we have presented a state-of-the-art subcategorisation acquisition system for free-word order languages, and used it to create a large subcategorisation frame lexicon for German verbs. Our SCF lexicon resource is available at `http://amor.cms.hu-berlin.de/˜robertsw/scflex.html`. We are performing a manual evaluation of the output of our system, which we will report soon.

We plan to continue this work first by expanding our SCF lexicon with case information and selectional preferences, second by using our SCF classifier and lexicon for verbal Multiword Expression identification in German, and last by comparing it to existing verb classifications, either by using available resources for German like the SALSA corpus (Burchardt et al., 2006), or by translating parts of VerbNet into German to create a more extensive gold standard for verb clustering in the spirit of Sun et al. (2010) who found that Levin's verb classification can be translated to French and still usefully allow generalisation over verb classes.

Finally, we plan to perform in vivo evaluation of our SCF lexicon, to determine what benefit it can deliver for NLP applications such as Semantic Role Labelling and Word Sense Disambiguation. Recent research has found that even automatically-acquired verb classifications can be useful for NLP applications (Shutova et al., 2010; Guo et al., 2011).

tion of the random cluster centroids used to initialise the k-means clustering). These samples are normally distributed, so we determine statistical significance using a t-test against the "Random Mean" results reported by Schulte im Walde (2006).

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.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *TLT*, pages 24–41.

Michael R. Brent and Robert C. Berwick. 1991. Automatic acquisition of subcategorization frames from tagged text. In *HLT*, pages 342–345. Morgan Kaufmann.

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. *CoRR*, cmp-lg/9702002.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2006. The SALSA corpus: A German corpus resource for lexical semantics. In *LREC*.

Judith Eckle-Kohler. 1999. *Linguistic knowledge for automatic lexicon acquisition from German text corpora*. Ph.D. thesis, Universität Stuttgart.

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.

Edward W. Forgy. 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, 21:768–769.

Raza Ghulam. 2011. *Subcategorization acquisition and classes of predication in Urdu*. Ph.D. thesis, Universität Konstanz.

Trond Grenager and Christopher D. Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *EMNLP*, pages 1–8.

Yufan Guo, Anna Korhonen, and Thierry Poibeau. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. In *EMNLP*, pages 273–283.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401.

Anna Korhonen. 2002. Subcategorization acquisition. Technical report, University of Cambridge, Computer Laboratory.

Joel Lang and Mirella Lapata. 2010. Unsupervised induction of semantic roles. In *HLT*, pages 939–947.

Lillian Lee. 1999. Measures of distributional similarity. In *ACL*, pages 25–32.

Beth Levin and Malka Rappaport Hovav. 2005. *Argument realization*. Cambridge University Press, Cambridge.

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

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *ACL*, pages 235–242.

Manolis Maragoudakis, Katia Lida Kermanidis, and George Kokkinakis. 2000. Learning subcategorization frames from corpora: A case study for modern Greek. In *Proceedings of COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries*, pages 19–22.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *ACL*, pages 433–440.

Anoop Sarkar and Daniel Zeman. 2000. Automatic extraction of subcategorization frames for Czech. In *COLING*, pages 691–697.

Silke Scheible, Sabine Schulte im Walde, Marion Weller, and Max Kisselew. 2013. A compact but linguistically detailed database for German verb subcategorisation relying on dependency parses from Web corpora: Tool, guidelines and resource. In *Web as Corpus Workshop*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *NeMLaP*, volume 12, pages 44–49.

Sabine Schulte im Walde and Chris Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *ACL*, pages 223–230.

Sabine Schulte im Walde. 2002a. A subcategorisation lexicon for German verbs induced from a lexicalised PCFG. In *LREC*, pages 1351–1357.

Sabine Schulte im Walde. 2002b. Evaluating verb subcategorisation frames learned by a German statistical grammar against manual definitions in the Duden Dictionary. In *EURALEX*, pages 187–197.

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 *COLING*, pages 1002–1010.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *ANLP*, pages 88–95.

Lin Sun, Anna Korhonen, and Yuval Krymolowski. 2008. Verb class discovery from rich syntactic data. In *CICLing*, pages 16–27, Haifa, Israel.

Lin Sun, Anna Korhonen, Thierry Poibeau, and Cédric Messiant. 2010. Investigating the cross-linguistic potential of VerbNet-style classification. In *COLING*, pages 1056–1064, Beijing, China.

Ivan Titov and Alexandre Klementiev. 2012. A Bayesian approach to unsupervised semantic role induction. In *EACL*, pages 12–22.

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

Oliver Wauschkuhn. 1999. *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora*. Shaker Verlag.