Evaluating Verb Subcategorisation Frames learned by a German Statistical Grammar against Manual Definitions in the *Duden* Dictionary

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
Azenbergstraße 12, 70174 Stuttgart, Germany
schulte@ims.uni-stuttgart.de

Abstract

The paper describes an extensive evaluation of computational large-scale verb subcategorisation by comparing subcategorisation frames induced from a German lexicalised statistical grammar against manual verb definitions in the dictionary *Duden - Das Stilwörterbuch*. We achieved an f-score of 62.30% on 3,090 verbs with a training corpus frequency between 10 and 2,000; ignoring prepositional phrases within the frame definitions resulted in an f-score of 72.05%. As to our knowledge, no former approach on automatic acquisition of verb subcategorisation has performed a comparably extensive evaluation. Our evaluation results justify the utilisation of the statistical grammar framework for obtaining a reliable subcategorisation lexicon for verbs. The lexical entries hold a potential for adding to and improving manual verb definitions.

1 Introduction

Subcategorisation properties of verbs represent an essential part of the verb lexicon; the verb itself is central to the meaning and the structure of a sentence, and lexical verb information represents the core in supporting NLP-tasks such as lexicography, parsing, machine translation, and information retrieval. Since manually built extensive lexica are resource-consuming, automatic subcategorisation lexica have been created, especially for English such as [Brent 1993; Manning 1993; Briscoe & Carroll 1997; Carroll & Rooth 1998]; and few for German such as [Eckle 1999; Wauschkuhn 1999]. In our approach, we obtained a large-scale computational subcategorisation lexicon by unsupervised learning in a statistical grammar framework [Schulte im Walde 2002]: a German context-free grammar containing frame- predicting grammar rules and information about lexical heads was trained on a large German newspaper corpus. The lexicalised version of the probabilistic grammar served as source for syntactic verb frame descriptions.

How reliable are such automatically created verb lexica? This paper describes the extensive evaluation of 3,090 verb entries within the learned subcategorisation lexicon against manual definitions in the German dictionary *Duden - Das Stilwörterbuch*. The work was performed in collaboration with *Bibliographisches Institut & F. A. Brockhaus AG* who provided a machine readable version of the dictionary.

As to our knowledge, no former approach on subcategorisation has performed a comparably extensive evaluation of computational large-scale verb subcategorisation. We show that (i) our evaluation results justify the utilisation of the statistical grammar framework for obtaining a reliable subcategorisation lexicon for verbs, and (ii) the lexical entries hold a potential for improving manual verb definitions.

2 Learning Verb Subcategorisation in a Statistical Grammar Framework

The large-scale computational subcategorisation lexicon was obtained from the trained parameters in a statistical grammar framework. Section 2.1 describes the grammar parameters relevant for the subcategorisation induction, and Section 2.2 illustrates the subcategorisation frame definition.

2.1 Statistical Grammar Framework

The acquisition of syntactic verb subcategorisation properties was performed by utilising the lexicalised probabilistic version of a German context-free grammar. The German grammar was developed with the goal of obtaining reliable lexical information on verbs. For example, the grammar contains a specific rule level

C -> S.<frame>

where the clause level C produces the clause category S accompanied by the relevant subcategorisation frame dominating the clause. The probabilistic version of the context-free grammar assigns frequencies to the grammar rules according to corpus appearance, to distinguish the relevance of different frame types:

freq_1 C -> S.<frame_1>
freq_2 C -> S.<frame_2>
$\text{freq}_{\dots}$ C -> S.<$\text{frame}_{\dots}$>
freq_n C -> S.<frame_n>

By that, we can make general statements about syntactic grammar structures, such as: *the transitive frame with a frequency of x is more frequent/probable than the expletive usage with a frequency of y.* But we are interested in the idiosyncratic, lexical usage of verbs, so we extend the probabilistic grammar by incorporating the lexical head of each rule into the grammar parameters

$\text{C}^{[\text{lex. head}]}$ -> S.<frame>

and the probabilistic version of the grammar rules distinguishes the relevance of different frame types according to a specific lexical head, i.e., the verb:

$\text{freq}_1{}^{[\text{lex. head}]}$ C -> S.<frame_1>
$\text{freq}_2{}^{[\text{lex. head}]}$ C -> S.<frame_2>
$\text{freq}_{\dots}{}^{[\text{lex. head}]}$ C -> S.<$\text{frame}_{\dots}$>
$\text{freq}_n{}^{[\text{lex. head}]}$ C -> S.<frame_n>

The simplified description of the above grammar rules describes the relevant grammar part for verb subcategorisation within *head-lexicalised probabilistic context-free grammars* (H-L PCFGs) [Carroll & Rooth 1998]. Statistical data on lexicalised grammar rules and lexical coherence parameters provide a basis for inducing lexical phenomena [Schulte im Walde et al. 2001]. To obtain the subcategorisation lexicon from the statistical grammar model, we performed unsupervised training on 18.7 million words of a large German newspaper corpus from the 1990s. The trained model served as lexical source for the large-scale computational acquisition of subcategorisation frames for 14,229 German verbs [Schulte im Walde 2002].

2.2 Subcategorisation Frame Definition

The subcategorisation frame types comprise maximally three arguments. Possible arguments in the frames are nominative (n), dative (d) and accusative (a) noun phrases, reflexive pronouns (r), prepositional phrases (p), expletive *es* (x), subordinated non-finite clauses (i), subordinated finite clauses (s-2 for verb second clauses, s-dass for *dass*-clauses, s-ob for *ob*-clauses, s-w for indirect *wh*-questions), and copula constructions (k). For example, subcategorising a direct (accusative) object (next to the obligatory (nominative) subject) would be represented by `na`; using an indirect (dative) object and a subcategorised non-finite clause by `ndi`. We defined a total of 38 subcategorisation frame types.

We used the trained frequency distributions over frame types for each verb (cf. Section 2.1) as basis for the subcategorisation properties of the respective verb. The frequency values were strengthened by squaring them. The strengthening enabled a clear-cut demarcation of lexically relevant and irrelevant frames, because the difference in frequencies was reinforced. The squared frequencies were normalised, and a cut-off of 1% defined the frames which are part of the lexical verb entry. Table 1 cites the (original and strengthened) frequencies and probabilities for the verb *zehren* 'to live on/wear down'; the table marks the demarcation of lexicon-relevant frames by an extra line in the columns on strengthened numbers.

Frame	***Freq (orig)***	***Prob (orig)***	***Freq^2 (strength)***	***Prob (strength)***
N	43	0.47110	1866	0.54826
np	39	0.42214	1499	0.44022
na	5	0.05224	23	0.00674
nap	4	0.04220	15	0.00440
nd	1	0.01232	1	0.00038

Lexical subcategorisation: {`n, np`}

Table 1: Probabilistic subcategorisation for *zehren*

A more delicate version of subcategorisation frames discriminates between the specific kinds of prepositional phrases for PP-arguments by distributing the frequency mass of prepositional phrase frame types (np, nap, ndp, npr, xp) over the prepositional phrases, according to their frequencies in the corpus, and setting a cut-off of 20%. Prepositional phrases are referred to by case and preposition, such as 'Dat.mit', 'Akk.für'. The resulting lexical subcategorisation for *zehren* would be {`n, np:Dat.von, np:Dat.an`}.

3 Manual Definition of Subcategorisation Frames in Dictionary *Duden*

The German dictionary *Duden - Das Stilwörterbuch* [Dudenredaktion 2001] describes the stylistic usage of words in sentences, such as their syntactic embedding, example sentences, idiomatic expressions. Part of the lexical verb entries are frame-like syntactic descriptions, such as `<von etw. zehren>` 'to live on something$_{Dat}$'. We extracted subcategorisation frames for 3,658 verbs from the *Duden*, with no restrictions concerning verb frequency or verb meaning.

Duden does not contain explicit subcategorisation frames, since it is not meant to be a subcategorisation lexicon. But for the description of the stylistic usage of verbs, the subcategorisation properties are a necessary element; therefore, the `grammatical information' contains implicit subcategorisation, which enables us to infer frame definitions. Alternations in verb meaning are marked by a semantic numbering and accompanied by the respective subcategorisation requirements. For example, the lexical verb entry for *zehren* lists the following lexical semantic verb entries:

1. `<von etw. zehren>` `live on something'
2. `drain somebody of his energy'
   a) no frame which implicitly refers to an intransitive usage
   b) `<an jmdm., etw. zehren>`

Idiosyncrasies in the manual frame definitions led to 1,221 different subcategorisation frames, e.g. identical frame definitions differ in their degree of explicitness, such as `<[gegen jmdn., etw. (Akk.)]>` and `<[gegen jmdn., etw.]>` which both refer to the potential subcategorisation of a prepositional phrase with accusative case and head *gegen* `against'. Correcting and reducing the frames resulted in 65 subcategorisation frame types.

4 Evaluation Experiments

Preceding the actual experiment I defined a mapping from the *Duden* frame definitions onto my subcategorisation frame style, e.g. the ditransitive frame definition `<jmdm. Etw.>` would be mapped to `nad`, `<bei jmdm. etw.>` to `nap` without or `nap:Dat.bei` with explicit prepositional phrase definition.

4.1 Recall, Precision, and F-Score Values

For the evaluation, the manual *Duden* frame definitions were considered as golden standard for the learned subcategorisation frames. We calculated precision and recall values on the following basis:

$$recall = tp / (tp + fn)$$
$$precision = tp / (tp + fp)$$

tp (true positives) refers to those subcategorisation frames where learned and manual definitions agree, *fn* (false negatives) to the *Duden* frames not filtered automatically, and *fp* (false positives) to those automatically filtered frames not defined by *Duden.*

Major importance was given to the f-score which considered recall and precision as equally relevant:

$$f\text{-}score = (2 * recall * precision) / (recall + precision)$$

4.2 Experiments

The experiment was three-fold.

I All frame types were taken into consideration. In case of a prepositional phrase argument in the frame, the PP was included, but the refined definition was neglected, e.g. the frame type including one obligatory prepositional phrase was referred to by `np` (nominative noun phrase plus prepositional phrase).

II All frame types were taken into consideration. In case of a prepositional phrase argument in the frame, the refined definition was included, e.g. the frame including one obligatory prepositional phrase (cf. I) was referred to by `np:Akk.für` for a prepositional phrase with head *für* and the accusative case, `np:Dat.bei` for a prepositional phrase with head *bei* and the dative case, etc.

III Prepositional phrases were excluded from subcategorisation, i.e. frames including a `p` were mapped to the same frame type without that argument. By that, a decision between prepositional phrase arguments and adjuncts was avoided.

4.3 Baseline

As baseline for the experiments, we assigned the most frequent frame types `n` (intransitive frame) and `na` (transitive frame) as default to each verb.

4.4 Results

Assuming that predictions on the most rare events (verbs with a low frequency) and on the most frequent verbs (with increasing tendency towards polysemy) are rather unreliable, we performed the evaluation on those 3,090 verbs with a frequency between 10 and 2,000. The results of the evaluation experiments are displayed in Table 2.

Experiment	*Recall*		*Precision*		*F-Score*	
	Baseline	Result	Baseline	Result	Baseline	Result
I	49.57%	63.91%	54.01%	60.76%	51.70%	62.30%
II	45.58%	50.83%	54.01%	65.52%	49.44%	57.24%
III	63.92%	69.74%	59.06%	74.53%	61.40%	72.05%

Table 2: Evaluation of subcategorisation frames

Concerning the f-score, we reached a gain of 10% compared to the baseline for experiment I: evaluating all frame definitions in the learned lexicon including prepositional phrases resulted in 62.30% f-score performance. Complicating the task by including prepositional phrase definitions into the frame types (experiment II), we reached 57.24% f-score performance, 8% above the baseline. Completely disregarding the prepositional phrases in the subcategorisation frames (experiment III) resulted in 72.05% f-score performance, 10% above the baseline.

The differences both in the absolute f-score values and the difference to the respective baseline values correspond to the difficulty and potential of the tasks. Disregarding the prepositional phrases completely (experiment III) is the easiest task and therefore reaches the highest f-score. But the baseline frames `n` and `na` represent 50% of all frames used in the *Duden* lexicon, so the potential for improving the baseline is small. Compared to experiment III, experiment I is a more difficult task, because the prepositional phrases are taken into account as well. But we could reach a gain in f-score of more than 10%, so the learned frames could improve the baseline decisions. Experiment II shows that defining prepositional phrases in verb subcategorisation is an even more complicated task. Still, we could improve the baseline results by 8%.

5 Lexicon Investigation

Section 4 presented the results of evaluating verb subcategorisation frames learned in a statistical grammar framework against the manual verb descriptions in the German dictionary *Duden*. The current section discusses advantages and shortcomings of the verb subcategorisation lexica concerning the selection of verbs and the set and detailness of frame types.

The verb entries in the automatic and manual subcategorisation lexica were investigated: the respective frames were compared, against each other as well as against verb entries in [Helbig & Schenkel 1969] (henceforth: H/S) and corpus evidence in the German newspaper corpus *die tageszeitung (TAZ)*. In addition, we compared the set of frames in the two lexica, their intersection and differences. The result of the investigation is a description of strengths and deficiencies in the lexica.

5.1 Intransitive Verbs

In the *Duden* dictionary, intransitive verb usage is difficult to filter, since it is defined only implicitly in the verb entry, such as for the verbs *glücken* `to succeed', *langen* `to suffice', *verzweifeln* `to despair'. In addition, *Duden* defines the intransitive frame for verbs which can be used intransitively in exclamations, such as *Der kann aber wetzen!* `Wow, he can dash!'. But the exclamatory usage is no sufficient evidence for intransitive usage. The learned lexicon, on the other hand, tends to overgenerate the intransitive usage of verbs, mainly because of parsing mistakes. Still, the intersection of intransitive frames in both lexica reaches a recall of 77.19% and a precision of 66.11%,

5.2 Transitive Verbs

The usage of transitive verbs in the lexica is the most frequent occurrence and at the same time the most successfully learned frame type. *Duden* defines transitive frames for 2,513 verbs, the automatic process filters 2,597 frames. An agreement in 2,215 cases corresponds to 88.14% recall and 85.29% precision.

5.3 Dative Constructions

Duden verb entries are inconsistent concerning the free dative construction (`freier Dativ'). For example, the free dative is existing in the ditransitive usage for the verb *ablösen* `to remove' (*Der Arzt löste ihm das Pflaster ab* `The doctor removed him the plaster'), but not for the verb *backen* `to bake' (H/S: *Die Mutter backt ihm einen Kuchen* `The mother baked him a cake'). The learned lexicon is rather unreliable on frames including dative noun phrases. Parsing mistakes tend to filter accusative constructions as dative and therefore wrongly emphasise the dative usage.

5.4 Prepositional Phrases

In general, *Duden* properly distinguishes between prepositional phrase arguments (mentioned in subcategorisation) and adjuncts, but in some cases, *Duden* overemphasises certain PP-arguments in the verb frame definition, such as `Dat.mit` for the verbs *aufschließen* `to unlock', *garnieren* `to garnish', *nachkommen* `to keep up', `Dat.von` for the verbs *abbröckeln* `to crumble', *ausleihen* `to borrow', *erbitten* `to ask for', *säubern* `to clean up', or **Akk.auf** for the verbs *abklopfen* `to check the reliability', *ausüben* `to practise', *festnageln* `to tie down', *passen* `to fit'.

In the learned lexicon, prepositional phrase arguments are overemphasised, i.e. PPs used as adjuncts are frequently inserted into the lexicon, such as for the verbs *arbeiten* `to work', *demonstrieren* `to demonstrate', *sterben* `to die'. This mistake is mainly based on highly frequent prepositional phrase adjuncts, such as **Dat.in, Dat.an, Akk.in**. On the other hand, the learned lexicon does not recognise verb-specific prepositional phrase arguments in some cases, such as **Dat.mit** for the verbs *gleichstellen* `to equate', *handeln* `to act', *spielen* `to play', or **Dat.von** for the verbs *abbringen* `to dissuade', *fegen* `to sweep', *genesen* `to convalesce', *schwärmen* `to romanticise'.

Comparing the frame definitions containing PPs in both lexica, the learned lexicon tends to define PP-adjuncts such as **Dat.in, Dat. an** as arguments and neglect PP-arguments; *Duden* distinguishes arguments and adjuncts more correctly, but tends to overemphasise PPs such as **Dat.mit** and **Dat.bei** as arguments. **np** frame agreement is still solved by 59.69% recall and 49.88% precision, but the evaluation of **nap** with 45.95% recall, 25.89% precision and of **ndp** with 9.52% recall and 15.87% precision pinpoints main deficiencies in the frame agreement.

5.5 Reflexive Verbs

Duden generously defines reflexive verbs; they appear whenever it is possible to use the respective verb with a reflexive pronoun. This idea is valid for verbs such as *erwärmen* `to heat', *lohnen* `to be worth', *schämen* `to feel ashamed', but overgenerating for verbs such as *durchbringen* `to pull through', *kühlen* `to cool', *zwingen* `to force'. The automatic frame definitions, on the other hand, tend to neglect the reflexive usage of verbs and rather choose direct objects into the frames, such as for the verbs *ablösen* `to remove', *erschießen* `to shoot', *überschätzen* `to overestimate'. The lexicon tendencies are reflected by the **nr, nar, npr** frame frequencies: rather low recall values between 28.74% and 45.17%, and rather high precision values between 51.94% and 69.34% underline the differences.

5.6 Adjectival Phrases

The definition of adjectival phrase arguments in the *Duden* is somewhat idiosyncratic, especially as demarcation to non-subcategorised adverbial phrases. For example, an adjectival phrase for the verb *scheinen* `to shine' as in *Die Sonne schien hell* `The sun is bright' is subcategorised, as well as for the verb *berühren* `to touch' as in *Seine Worte haben uns tief berührt* `His words touched us deeply'. Concerning the learned lexicon, the grammar does not contain adjectival phrase arguments, so they could not be recognised, such as for the verbs *anmuten* `to seem', *erscheinen* `to seem', *verkaufen* `to sell'.

5.7 Subcategorisation of Clauses

Duden shows shortcomings on the subcategorisation of non-finite and finite clauses; they rarely appear in the lexicon. Only 26 verbs (such as *anweisen* `to instruct', *beschwören* `to swear', *versprechen* `to promise') subcategorise non-finite clauses, only five verbs (such as *sehen* `to see', *wundern* `to wonder') subcategorise finite clauses. Missing verbs for the subcategorisation of finite clauses are -among others- *ausschließen* `to rule out', *sagen* `to say', *vermuten* `to assume', for the subcategorisation of non-finite clauses *hindern* `to prevent', *verpflichten* `to commit'.

The automatic lexicon defines the subcategorisation of clauses more reliably. For example, the verbs *behaupten* 'to state', *nörgeln* 'to grumble' subcategorise verb second finite clauses, the verbs *aufpassen* 'to pay attention', *glauben* 'to think', *hoffen* 'to hope' subcategorise finite *dass*-clauses, the verb *bezweifeln* 'to doubt' subcategorises a finite *ob*-clause, the verbs *ahnen* 'to guess', *klarmachen* 'to make clear', *raffen* 'to understand' subcategorise indirect *wh*-questions, and the verbs *anleiten* 'to instruct', *beschuldigen* 'to accuse', *lehren* 'to teach' subcategorise non-finite clauses. Mistakes occur for indirect *wh*-questions which are confused with relative clauses, such as for the verbs *ausbaden* 'to pay for', *futtern* 'to eat'.

5.8 General Frame Description

Duden defines verb usage on various levels of detailness, especially concerning prepositional phrases (cf. Section 2.2). For example, irgendwie 'somehow' in grammatical definitions means the usage of either a prepositional phrase such as for the verb *lagern* 'to store' (*Medikamente müssen im Schrank lagern* 'Drugs need to be stored in a cupboard'); irgendwo 'somewhere' means the usage of a locative prepositional phrase such as for the verb *lauern* 'to lurk' (*Der Libero lauert am Strafraum* 'The sweeper lies in wait in the penalty area'). In more restricted cases, the explicit prepositional phrase is given as in <über etw. (Akk.)> for the verb *verzweifeln* 'to despair' (*Man könnte verzweifeln über so viel Ignoranz* 'One could despair about that ignorance').

The grammatical definitions on various levels of detailness are considered as a strength of *Duden* and generally favourable for users of a stylistic dictionary, but produce difficulties for automatic usage. For example, when including PP-definitions into the evaluation (experiment II), 10% of the *Duden* frames (PP frames without explicit PP-definition, such as np) could never be guessed correctly, since the automatic lexicon includes the PPs explicitly.

There are frame types in *Duden* which do not exist in the automatic verb lexicon. This mainly concerns rare frames such as nag, naa, xad and frame types with more than three arguments such as napr, ndpp. This lexicon deficiency concerns about 4% of the total number of frames in the *Duden* lexicon.

5.9 Lexicon Coverage

Compared to the automatic acquisition of verbs, *Duden* misses verbs in the dictionary: frequent verbs such as *einreisen* 'to enter', *finanzieren* 'to finance', *veranschaulichen* 'to illustrate', verbs adopted from English such as *dancen, outen, tunen,* vulgar verbs such as *anpöbeln* 'to abuse', *ankotzen* 'to make sick', *pissen* 'to piss', recent neologisms such as *digitalisieren* 'to digitalise', *klonen* 'to clone', and regional expressions such as *kicken* 'to kick', *latschen* 'to walk', *puhlen* 'to pick'.

The automatic acquisition of verbs covers a larger amount of verbs, containing 14,229 verb entries, including the missing examples above. Partly, mistaken verbs are included in the lexicon: verbs wrongly created by the morphology such as **angebieten, *dortdrohen, *einkommen*, verbs which obey the old, but not the reformed German spelling rules such as *autofahren* 'to drive a car', *danksagen* 'to thank', *spazierengehen* 'to stroll', and rare verbs, such as $^{?}$*bürgermeistern*, $^{?}$*evangelisieren*, $^{?}$*fiktionalisieren*, $^{?}$*feuerwerken*, $^{?}$*käsen*.

5.10 Summary

Table 3 summarises the lexicon investigation. We blindly classified 184 frame assignments from *fn* and *fp* into correct and wrong. The result emphasises (i) unreliabilities for n and nd in both lexica, (ii) insecurities for reflexive and expletive usage in both lexica, (iii) strength of clause subcategorisation in the learned lexicon (the few assignment in *Duden* were all correct), (iv) strength of PP-assignment in the *Duden*, and (v) variability of PP-assignment in the learned lexicon.

The lexicon investigation showed that

- in both lexica, the degree of reliability of verb subcategorisation information depends on the different frame types,
- we need to distinguish between the different goals of the subcategorisation lexica: the learned lexicon explicitly refers to verb arguments which are (obligatorily) subcategorised by the verbs in the lexicon, whereas *Duden* was not intended to represent a subcategorisation lexicon but rather describe the stylistic usage of the verbs and therefore refer to possibly subcategorised verb arguments; in the latter case, there is no distinction between obligatory and possible verb complementation,
- a manual lexicon suffers from the human potential of permanently establishing new words in the vocabulary; it is difficult to be up-to-date, and the learned lexical entries therefore hold a potential for adding to and improving manual verb definitions.

Frame Type	***Duden: fn***		***Learned: fp***	
	correct	wrong	correct	wrong
`n`	4	6	3	7
`nd`	2	8	0	10
`nr, nar, ndr`	5	5	3	7
`x, xa, xd, xr`	6	4	3	7
`ni, nai, ndi`			5	5
`ns/nas/nds-dass`			9	0
`ns/nas/nds-2`			9	1
`np/nap/ndp/npr:Dat.mit`	7	3	6	4
`np/nap/ndp/npr:Dat.von`	7	3	5	0
`np/nap/ndp/npr:Dat.in`	6	4	3	7
`np/nap/ndp/npr:Dat.an`	9	1	6	4

Table 3: Lexicon investigation on *fn* and *fp*

6 Related Work

As to our knowledge, no former approach on subcategorisation has performed a comparably extensive evaluation of computational large-scale verb subcategorisation. Concerning subcategorisation lexica for English, [Brent 1993] evaluated learned subcategorisation frames against hand judgements. Results are recall of 60.00% and precision of 96.00%, which corresponds to an f-score of 73.85%. Differently to the following approaches, the number of frame types was restricted to six. In addition, the frames did not include prepositional phrase definitions. [Manning 1993] randomly selected 40 verbs from a list of 2,000 common verbs and evaluated learned subcategorisation frames (including prepositional phrase definitions) against *The Oxford Advanced Learner's Dictionary*. The results were recall of 43.00% and precision of 90.00%, which corresponds to an f-score of 58.20%. [Briscoe & Carroll 1997] performed an evaluation of learned subcategorisation frames (including prepositional phrase definitions) against the *Alvey NL Tools* dictionary [Boguraev et al. 1987] and the *COMLEX Syntax* dictionary [Grishman et al. 1994]. The evaluation was only on 14 verbs, resulting in recall of 65.70%, precision of 35.50% and f-score of 46.09%. [Carroll & Rooth 1998] utilised *The Oxford Advanced Learner's Dictionary* for evaluating learned subcategorisation frames for 200 randomly chosen verbs with a frequency greater than 500. The frames did not include prepositional phrase definitions. Results are recall of 75.00% and precision of 79.00%, which corresponds to an f-score of 76.95%.

Concerning lexica for German, [Eckle 1999] evaluated her subcategorisation frames (including prepositional phrases) on only 15 verbs against *Duden - Das große Wörterbuch der deutschen Sprache* [Drowdowski 1993]. She does not cite explicit recall and precision values, except for a subset of subcategorisation frames. [Wauschkuhn 1999] chose seven verbs with various subcategorisation frames (including prepositional phrases) out of 1,044 verbs in his automatic acquisition approach. He evaluated against hand judgement and achieved recall of 56.60% and precision of 68.20%, which corresponds to an f-score of 61.86%.

None of the approaches -neither for English nor for German verbs- considered more than 200 verbs for the evaluation of subcategorisation frames. The most successful subcategorisation definition (disregarding prepositional phrase definitions) took place in [Carroll & Rooth 1998]. But their evaluation was facilitated by restricting the frequency of the evaluated verbs to more than 500. [Brent 1993] outperformed our f-score result, but he did only use five frame types. [Manning 1993] and [Briscoe & Carroll 1997] are closely related in their evaluation of subcategorisation to our approach. They also evaluated frame types including prepositional phrases against dictionaries and reached f-scores of 58.20% and 46.09%, respectively, compared to our result of 57.24%.

There are no directly comparable evaluations for German, since both German approaches on learning verb subcategorisation evaluated on a hand-selected, low number of verbs.

7 Summary

We performed an extensive evaluation of computational large-scale verb subcategorisation by comparing verb subcategorisation frames learned by a German statistical grammar against manual verb entries in *Duden - Das Stilwörterbuch*. We achieved an f-score of 62.30% (10% above the baseline) on 3,090 verbs with a training corpus frequency between 10 and 2,000. Ignoring prepositional phrases within the frame definitions resulted in an f-score of 72.05% (10% above the baseline), specifying the prepositional phrases within the frame definitions by case and prepositional head resulted in an f-score of 57.24% (8% above the baseline). The differences in the results emphasise the particular difficulty of distinguishing between prepositional phrase arguments and adjuncts.

As to our knowledge, no former approach on subcategorisation has performed a comparably extensive evaluation of computational large-scale verb subcategorisation. Existing evaluations for English considered either less verbs or restricted the frequencies of the evaluated verbs. For German, learned subcategorisation frames were evaluated only on a hand-selected, low number of verbs.

Our evaluation results justify the utilisation of the statistical grammar framework for obtaining a reliable subcategorisation lexicon for verbs. Large-scale computational subcategorisation properties for several thousand verbs are provided, unrestricted concerning the verb frequencies, referring to the diversity of text genre given in newspaper corpora. The lexical entries hold a potential for adding to and improving manual verb definitions.

References

[Boguraev et al. 1987] Boguraev, B., Briscoe, E., Carroll, J., Carter, D., and Grover, C., 1987. The Derivation of a Grammatically-Indexed Lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 193-200, Stanford, CA.

[Brent 1993] Brent, M.R., 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19:203-222.

[Briscoe & Carroll 1997] Briscoe, T. and Carroll, J., 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing.*

[Carroll & Rooth, 1998] Carroll, G. and Rooth, M., 1998. Valence Induction with a Head-Lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain.

[Drosdowski 1993] Drosdowski, G., 1993. *Das große Wörterbuch der deutschen Sprache.* Dudenverlag, Mannheim.

[Dudenredaktion 2001] Dudenredaktion, editor, 2001. *Duden - Das Stilwörterbuch.* Number 2 in 'Duden in zwölf Bänden'. Dudenverlag, Mannheim, 8th edition.

[Eckle 1999] Eckle, J., 1999. *Linguistic Knowledge for Automatic Lexicon Acquisition from German Text Corpora.* PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[Grishman et al. 1994] Grishman, R., Macleod, C., and Meyers, A., 1994. Comlex Syntax: Building a Computational Lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 268-272, Kyoto, Japan.

[Helbig & Schenkel 1969] Helbig, G. and Schenkel, W., 1969. *Wörterbuch zur Valenz und Distribution deutscher Verben.* Max Niemeyer Verlag, Tübingen.

[Manning 1993] Manning, C.D., 1993. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics,* pages 235-242.

[Schulte im Walde 2002] Schulte im Walde, S., to appear. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings of the 3rd Conference on Language Resources and Evaluation*, Las Palmas, Spain.

[Schulte im Walde et al. 2001] Schulte im Walde, S., Schmid, H., Rooth, M., Riezler, S., and Prescher, D., 2001. Statistical Grammar Models and Lexicon Acquisition. In Rohrer, C., Rossdeutscher, A., and Kamp, H., editors, *Linguistic Form and its Computation.* CSLI Publications, Stanford, CA.

[Wauschkuhn 1999] Wauschkuhn, O., 1999. *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora. PhD thesis*, Institut für Informatik, Universität Stuttgart.