A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG

Sabine Schulte im Walde

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

Abstract

The paper presents a large-scale computational subcategorisation lexicon for several thousand German verbs. The lexical entries were obtained by unsupervised learning in a statistical grammar framework: a German context-free grammar containing frame-predicting grammar rules and information about lexical heads was trained on 18.7 million words of a large German newspaper corpus. We developed a simple methodology to utilise frequency distributions in the lexicalised version of the probabilistic grammar for inducing syntactic verb frame descriptions. The frame definition is variable with respect to the inclusion of prepositional phrase refinement. An evaluation against a manual dictionary justifies the utilisation of the machine-readable lexicon as a valuable component for supporting NLP-tasks. As to our knowledge, no former computational approach has obtained a subcategorisation lexicon for German comparable in size (the number of verbs in the lexicon), restriction (no limit concerning the frequencies of the verbs), or verified reliability (successful extensive evaluation against dictionary).

1. Introduction

Subcategorisation properties of verbs constitute 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 and Carroll, 1997; Carroll and Rooth, 1998), few for German such as (Eckle, 1999; Wauschkuhn, 1999).

We present a large-scale computational subcategorisation lexicon for several thousand German verbs, unrestricted concerning the verb frequencies. The lexical entries are induced from a lexicalised probabilistic context-free grammar: we developed a German context-free grammar containing frame-predicting grammar rules, refined the grammar by lexical head information and trained unsupervised within a statistical framework on lexicon induction (Schulte im Walde et al., 2001). The resulting statistical grammar model serves as source for lexical information: a simple methodology utilises frequency distributions in the lexicalised grammar for inducing syntactic verb descriptions. The frame definition is variable with respect to the inclusion of prepositional phrase refinement.

The induced subcategorisation information (i) constitutes lexical verb entries with clear demarcation of lexically relevant frame definitions, and (ii) provides frequency and probability distributions over frame types, for finer-grained usage in lexicon-related NLP-tasks. An evaluation against a manual dictionary shows that the lexical entries hold a potential for adding to and improving manual verb definitions.

2. Grammar Development

We developed a context-free grammar for German, with the goal of obtaining reliable lexical information on verbs. Work effort therefore concentrated on defining linguistic structures which are relevant to lexical verb information, especially subcategorisation.

On the one hand, this resulted in fine-grained structural levels for subcategorisation. The following paragraphs will illustrate the verb-related grammar structure.

The grammar distinguishes six finite clause types,

- `C-1-2` for verb first and verb second clauses,
- `C-rel` for relative clauses,
- `C-sub` for non-subcategorised subordinated clauses,
- `C-dass` for subcategorised subordinated *dass*-clauses,
- `C-ob` for subcategorised subordinated *ob*-clauses,
- `C-w` for subcategorised indirect *wh*-questions.

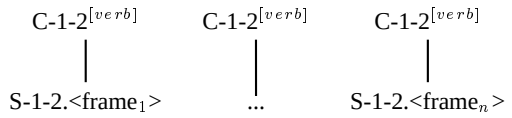
For each clause type, I introduced an extraordinary rule level

`C-<type>` $\rightarrow$ `S-<type>.<frame>`

where the clause level `C` produces the clause category `S` which is accompanied by the relevant subcategorisation frame dominating the clause. A lexicalisation of the grammar rules with their verbal heads automatically leads to a distribution over frame types.

`C-<type>`$^{[verb]}$ $\rightarrow$ `S-<type>.<frame>`

Therefore, by introducing the extra clause level `C`, a specific level for frame selection was created:



Frame Type	Example
n	*Sie$_n$ schwimmt.*
na	*Er$_n$ sieht sie$_a$.*
nd	*Er$_n$ glaubt ihr$_d$.*
np	*Sie$_n$ achten auf Kinder$_p$.*
nad	*Sie$_n$ verspricht ihm$_d$ ein Geschenk$_a$.*
nap	*Sie$_n$ hindert ihn$_a$ am Stehlen$_p$.*
ndp	*Er$_n$ dankt ihr$_d$ für ihr Verständnis$_p$.*
ni	*Er$_n$ versucht, pünktlich zu kommen$_i$.*
nai	*Er$_n$ hört sie$_a$ ein Lied singen$_i$.*
ndi	*Sie$_n$ verspricht ihm$_d$ zu kommen$_i$.*
nr	*Sie$_n$ fürchten sich$_r$.*
nar	*Er$_n$ erhofft sich$_r$ Aufwind$_a$.*
ndr	*Sie$_n$ schließt sich$_r$ der Kirche$_d$ an.*
npr	*Er$_n$ hat sich$_r$ als würdig$_p$ erwiesen.*
nir	*Sie$_n$ stellt sich$_r$ vor, alles zu gewinnen$_i$.*
x	*Es$_x$ blitzt.*
xa	*Es$_x$ gibt viele Bücher$_a$.*
xd	*Es$_x$ graut mir$_d$.*
xp	*Es$_x$ geht um ein tolles Angebot$_p$.*
xr	*Es$_x$ rechnet sich$_r$.*
xs-dass	*Es$_x$ heißt, dass er sehr klug ist$_{s-dass}$.*
ns-2	*Er$_n$ hat gesagt, er halte einen Vortrag$_{s-2}$.*
nas-2	*Er$_n$ schnauzt ihn$_a$ an, er sei ein Idiot$_{s-2}$.*
nds-2	*Er$_n$ sagt ihr$_d$, sie sei unmöglich$_{s-2}$.*
nrs-2	*Er$_n$ wünscht sich$_r$, sie bliebe bei ihm$_{s-2}$.*
ns-dass	*Er$_n$ hat angekündigt, dass er kommt$_{s-dass}$.*
nas-dass	*Er$_n$ fordert sie$_a$ auf, dass sie verreist$_{s-dass}$.*
nds-dass	*Er$_n$ sagt ihr$_d$, dass er unmöglich sei$_{s-dass}$.*
nrs-dass	*Er$_n$ wünscht sich$_r$, dass sie bleibt$_{s-dass}$.*
ns-ob	*Er$_n$ hat gefragt, ob sie den Vortrag hält$_{s-ob}$.*
nas-ob	*Er$_n$ fragt sie$_a$, ob sie ihn liebt$_{s-ob}$.*
nds-ob	*Er$_n$ ruft ihr$_d$ zu, ob sie verreist$_{s-ob}$.*
nrs-ob	*Er$_n$ wird sich$_r$ erinnern, ob sie dort war$_{s-ob}$.*
ns-w	*Er$_n$ hat gefragt, wann sie ankommt$_{s-w}$.*
nas-w	*Er$_n$ fragt sie$_a$, warum sie ihn liebt$_{s-w}$.*
nds-w	*Er$_n$ sagt ihr$_d$, wer zu Besuch kommt$_{s-w}$.*
nrs-w	*Er$_n$ erinnert sich$_r$, wer zu Besuch kommt$_{s-w}$.*
k	*Er ist ein Idiot$_k$.*

Table 1: Subcategorisation frame types

Abstracting from the clause type, the combination of grammar rules and lexical (verb) head information provides distributions for each verb over its subcategorisation frame properties.

The clause category S produces verb phrases; they are defined as verb complexes which collect preceding and following arguments and adjuncts until the sentence is parsed. The resulting frame indicates the number and types of the verbal 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). The resulting 38 frame types are listed in Table 1, accompanied by verb second example clauses.

The parsing strategy is organised in an exceptional way: we are interested in the head information on clause level.

Clause Type	Example
verb first clause	*Liebt* Peter seine Freundin? Hat Peter seine Freundin *geliebt*?
verb second clause	Peter *liebt* seine Freundin. Peter hat seine Freundin *geliebt*.
verb final clause	weil Peter seine Freundin *liebt* weil Peter seine Freundin *geliebt* hat
relative clause	der seine Freundin *liebt* der seine Freundin *geliebt* hat
indirect *wh*-question	wer seine Freundin *liebt* wer seine Freundin *geliebt* hat
non-finite clause	seine Freundin zu *lieben* seine Freundin *geliebt* zu haben

Table 2: Clause type examples

Since the verbal lexical head as the bearer of the clausal subcategorisation needs to be propagated through the parse tree, the grammar structures are based on a so-called 'collecting strategy' around the verbal head. The collection of verbal adjacents is performed differently according to the clause type, since the relevant lexical verbal head may be realised by different syntactic categories in different positions, cf. example sentences in Table 2.

As examples for different clausal parses, Figure 1 shows the syntactic tree analysis for the ditransitive verb second clause *Er schenkt seiner Freundin Schokolade* 'he gives chocolate to his girl-friend' (where it is necessary to collect all but one argument to the right of the finite verb), Figure 2 for the respective clause in a verb-final construction: *weil er seiner Freundin Schokolade schenkt* 'because he gives chocolate to his girl-friend' (where it is necessary to collect all arguments to the left of the finite verb). The verb phrase annotation indicates the clause types (1-2 and F) for the verb phrases, the frame type nad and the yet collected arguments (_ for none). The lexical heads of the nodes in the tree are marked by superscripts of the syntactic categories.

Noun phrases in the grammar are represented by NP plus case indication such as NP.Nom, prepositional phrases by PP plus case and prepositional head indication such as PP.Dat.mit. Structural levels for constituents outside verbal subcategorisation are disregarded. For example, adjectival and adverbial phrases are defined by a simple bar-structure which is able to recognise the phrases reliably, but disregards a fine-tuning of their internal structure.

The German context-free grammar consists of 35,821 rules, with 94% of them modelling verb subcategorisation.

3. Grammar Training

The context-free grammar was set into a head-lexicalised probabilistic environment by incorporating the lexical head of each rule into the grammar parameters and assigning probabilities to the rules. Grammar training was then performed by the statistical parser LoPar (Schmid, 2000).

Head-lexicalised probabilistic context-free grammars (H-L PCFGs) are a lexicalised extension of PCFGs; they incorporate the lexical head of each rule into the grammar parameters. The definition originated in (Carroll and Rooth, 1998) and was refined by (Schmid, 2000). According to

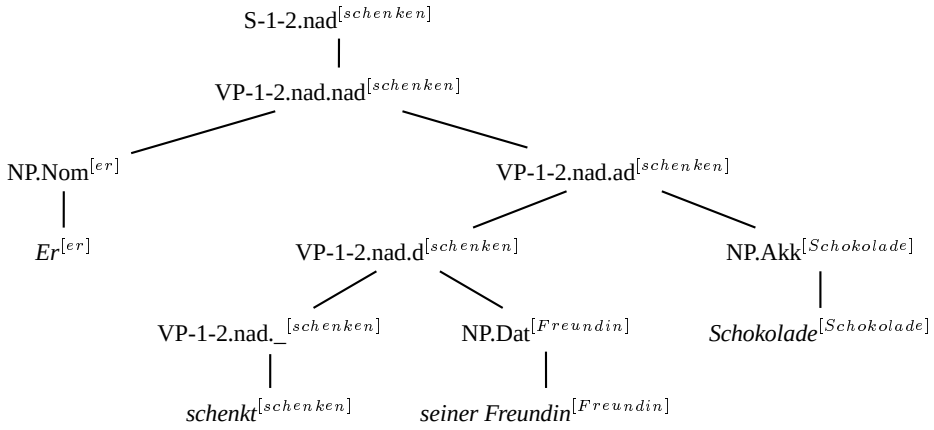


Figure 1: Syntactic analysis for verb second

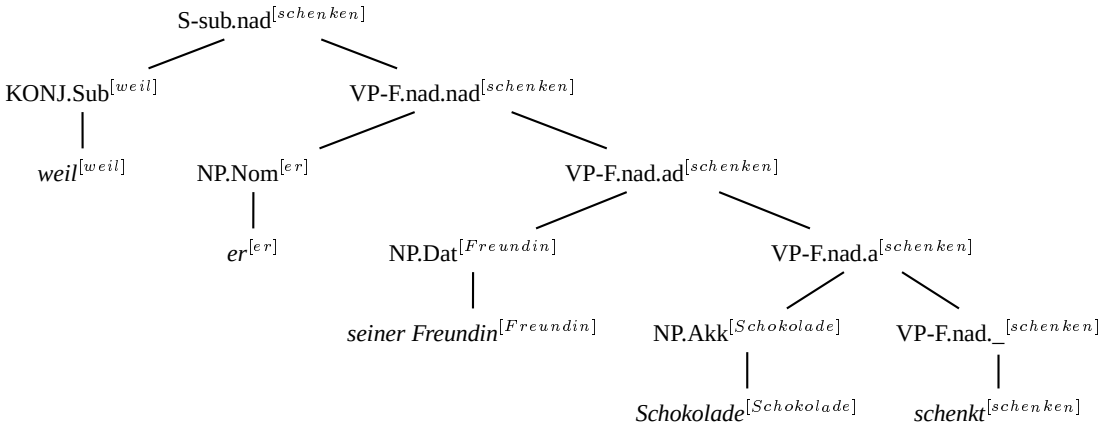


Figure 2: Syntactic analysis for verb final

an H-L PCFG, the probability of a syntactic tree analysis $p(T)$ for a sentence is defined as the product of the probabilities for choosing the start category s_j, the rules r_i, and the relevant lexical heads h which are included in the tree.

$$
\begin{aligned}
p(T) \quad = \quad & p_{start}(s_j) \ * \\
& p_{start}(h|s_j) \ * \\
& \prod_{r_i \in R, C_P \in N, h \in T} p_{rule}(r_i|C_P,h)^{f_r} \ * \\
& \prod_{C_P, C_C \in N, h_P, h_C \in T} p_{choice}(h_C|C_P,h_P,C_C)^{f_{ch}}
\end{aligned}
$$

$p_{start}(s_j)$ is the probability that the start category s_j is the category of the root node of a parse tree. $p_{start}(h|s_j)$ is the probability that a root node of category s_j bears the lexical head h. $p_{rule}(r_i|C_P,h)$ is the probability that a (parent) node of the category C_P with lexical head h is expanded by the grammar rule r_i. $p_{choice}(h_C|C_P,h_P,C_C)$ is the probability that a (non-head) child node of category C_C bears the lexical head h_C given that the parent category is C_P and the parent head is h_P. R refers to the set of rules established by the grammar, N to the set of non-terminal categories, and T to the set of terminal categories. Frequencies in the tree analysis are referred to by $f_r = f_{Tree}(r_i, C_P, h)$ for lexical rule parameters and $f_{ch} = f_{Tree}(h_C, C_P, h_P, C_C)$ for lexical choice parameters.

Head-lexicalised PCFGs enable to train and use grammar models which enrich purely syntactic information with lexical specification. On the basis of lexicalised probabilities, H-L PCFGs are able to rank syntactic analyses with reference to lexical preferences.

The statistical parser `LoPar` is an implementation of the left-corner parsing algorithm. Its functionality comprises symbolic parsing with context-free grammars and probabilistic training and parsing with probabilistic context-free grammars (PCFGs) and head-lexicalised probabilistic context-free grammars (H-L PCFGs). In addition, the parser can be applied for Viterbi parsing, tagging and chunking. `LoPar` executes the parameter training of the extended context-free grammars by the unsupervised *Inside-Outside Algorithm* (Lari and Young, 1990), an instance of the *Expectation-Maximisation Algorithm* (Baum, 1972). The algorithm iteratively improves model parameters by alternately assessing frequencies and estimating probabilities.

For inducing the subcategorisation lexicon from the statistical grammar model, we performed unsupervised training within five training iterations on 18.7 million words of a large German newspaper corpus from the 1990s.

4. Lexicon Induction

The trained H-L PCFG served as source for the computational induction of subcategorisation frames for lexical

verb entries. The statistical grammar model contains lexicalised subcategorisation rules for 14,229 verbs with a frequency between 1 and 255,676.

Recall the grammar rules for verb subcategorisation

`C-<type>` $\rightarrow$ `S-<type>.<frame>`

with `S` accompanied by the subcategorisation frames. The lexicalised version of the probabilistic grammar combines the set of rules with their lexical heads:

`C-<type>`$^{[verb]}$ $\rightarrow$ `S-<type>.<frame>`

The lexicalised rule provides a frequency distribution for verbs over subcategorisation frame types. For example, Table 3 presents the respective lexicalised rule parameters (in verb second clauses) for the verb *glauben* 'to think/believe'.

Freq	Grammar Rule	Lex. Head
1,921	C-1-2 $\rightarrow$ S-1-2.ns-dass	glauben
1,880	C-1-2 $\rightarrow$ S-1-2.ns-2	glauben
687	C-1-2 $\rightarrow$ S-1-2.np	glauben
498	C-1-2 $\rightarrow$ S-1-2.na	glauben
423	C-1-2 $\rightarrow$ S-1-2.n	glauben
341	C-1-2 $\rightarrow$ S-1-2.ni	glauben
210	C-1-2 $\rightarrow$ S-1-2.nd	glauben
144	C-1-2 $\rightarrow$ S-1-2.nad	glauben
69	C-1-2 $\rightarrow$ S-1-2.nds-2	glauben
57	C-1-2 $\rightarrow$ S-1-2.ns-w	glauben
49	C-1-2 $\rightarrow$ S-1-2.nai	glauben
46	C-1-2 $\rightarrow$ S-1-2.nas-w	glauben
36	C-1-2 $\rightarrow$ S-1-2.nap	glauben
29	C-1-2 $\rightarrow$ S-1-2.nar	glauben
27	C-1-2 $\rightarrow$ S-1-2.nrs-2	glauben
27	C-1-2 $\rightarrow$ S-1-2.ndp	glauben
24	C-1-2 $\rightarrow$ S-1-2.nr	glauben
20	C-1-2 $\rightarrow$ S-1-2.nas-dass	glauben
18	C-1-2 $\rightarrow$ S-1-2.npr	glauben
17	C-1-2 $\rightarrow$ S-1-2.nds-dass	glauben
14	C-1-2 $\rightarrow$ S-1-2.nas-2	glauben
10	C-1-2 $\rightarrow$ S-1-2.ndi	glauben

Table 3: Lexicalised rules for subcategorisation

Abstracting from the clause type, we used the trained frequency distributions over frame types for each verb 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 4 cites the (original and strengthened) frequencies and probabilities for the verb *glauben*, Table 5 for the verb *zehren* 'to live on/wear down'; each table marks the demarcation of lexicon-relevant frames by an extra line in the rows on strengthened numbers.

We also created a more delicate version of subcategorisation frames that discriminates between different kinds of PP-arguments. This was done 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. Prepositional phrases are referred to by case and preposition, such as 'Dat.mit', 'Akk.für'.

As for the subcategorisation frame types, we could filter the usage of prepositional phrases from the lexicalised rule parameters. For example, the parameter

`96 VP.np.np`$^{[reden]}$ $\rightarrow$ `VP.np.n' PP.Akk.über`

determines that `PP.Akk.über` with accusative case and prepositional head *über* is subcategorised (notice the change in frame saturation between parent and child `VP`) 96 times by the lexical verb head *reden* 'to talk' on its right hand side, within the frame type `np`. Abstracting from the position of the prepositional phrase and summing over the respective rule frequencies results in a frequency distribution over prepositional phrase types within subcategorisation frames. Table 6 shows the frequency distribution for the verb *reden* and the frame type `np` (with frequencies $\geq$ 10).

PP Type		Freq
Akk.über	acc / 'about'	480
Dat.von	dat / 'about'	463
Dat.mit	dat / 'with'	280
Dat.in	dat / 'in'	81
Nom.vgl	nom / 'as'	14
Dat.bei	dat / 'at'	13
Dat.über	dat / 'about'	13
Dat.an	dat / 'at'	12
Akk.für	acc / 'for'	10

Table 6: PP frequencies for *reden* and `np`

The frequency values for the prepositional phrases were also strengthened by squaring them, and the squared frequencies were normalised. When refining subcategorisation frames by PPs, the joint probability of the verb and the respective frame type (e.g. *reden* subcategorises the frame type `np` with a joint probability of 0.35820) was distributed over the different kinds of PPs, according to the probability of the PP type given the PP frame type. If the probability product exceeded a cut-off of 20% –which lies strong restrictions on the usage of PPs as arguments– the joint combination of frame type and PP (e.g. `np:Akk.über`) was marked as lexicon-relevant.

5. Lexicon Representation

We created a subcategorisation lexicon for 14,229 German verbs. The lexicon database contains the verb lemma, the frequency (according to the training corpus), and a list of those subcategorisation frames which were considered to be lexicon-relevant, (i) for the basic frame types, and (ii) for the frame types refined by prepositional phrases. Table 7 lists examples for lexical entries without prepositional phrase refinement, Table 8 lists examples for lexical entries including the PP refinement.

The lexicon constitutes lexical verb entries with clear demarcation of lexically relevant frame definitions; alternatively, NLP tasks can utilise the finer-grained lexical verb subcategorisation information, i.e. the frequency and probability distributions of verbs over frame types.

Frame	Freq (orig)	Prob (orig)	Freq^2 (strength)	Prob (strength)
ns-dass	1,921	0.29283	3,688,398	0.44328
ns-2	1,880	0.28668	3,535,077	0.42485
np	687	0.10469	471,433	0.05666
na	498	0.07588	247,626	0.02976
n	423	0.06444	178,625	0.02147
ni	341	0.05201	116,336	0.01398
nd	210	0.03209	44,285	0.00532
nad	144	0.02201	20,846	0.00251
nds-2	69	0.01057	4,807	0.00058
ns-w	57	0.00874	3,282	0.00039
nai	49	0.00747	2,402	0.00029
nas-w	46	0.00702	2,120	0.00025
nap	36	0.00547	1,287	0.00015
nar	29	0.00443	843	0.00010
nrs-2	27	0.00413	734	0.00009
ndp	27	0.00407	711	0.00009
nr	24	0.00359	554	0.00007
nas-dass	20	0.00304	397	0.00005
npr	18	0.00274	324	0.00004
nds-dass	17	0.00263	297	0.00004
nas-2	14	0.00215	200	0.00002
ndi	10	0.00154	102	0.00001

Lexical subcategorisation: { ns-dass, ns-2, np, na, n, ni }

Table 4: Probabilistic subcategorisation for *glauben*

Frame	Freq (orig)	Prob (orig)	Freq^2 (strength)	Prob (strength)
n	43	0.47110	1,866	0.54826
np	39	0.42214	1,499	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 5: Probabilistic subcategorisation for *zehren*

6. Evaluation

The subcategorisation lexicon was evaluated against manual definitions in the German dictionary *Duden – Das Stilwörterbuch* (Dudenredaktion, 2001; Schulte im Walde, 2002). The evaluation was based on an extensive choice of 3,090 verbs from the automatic lexicon, with a verb frequency between 10 and 2,000 and no restrictions concerning the verb meaning. We calculated precision and recall values on the following basis:

$$recall = \frac{tp}{tp + fn} \quad (1)$$

$$precision = \frac{tp}{tp + fp} \quad (2)$$

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:

$$f - score = \frac{2 * recall * precision}{recall + precision} \quad (3)$$

We achieved an f-score of 62.30% (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).

Shortcomings in the automatic lexicon mainly concerned intransitive and dative constructions as well as the distinction between prepositional phrase arguments and adjuncts. Strength was particularly attributed to the subcategorisation of finite and non-finite clauses. Partly, mistaken verbs were included in the lexicon: verbs wrongly created by the morphology such as **angebieten*, **dortdrohen*, **einkommen*, and verbs which obey the old, but not the reformed German spelling rules such as *autofahren*, *danksagen*, *spazierengehen*.

Lexicon Entry			
Verb		Freq	Subcategorisation
aufregen	'to get excited'	135	na, nr
beauftragen	'to order', 'to charge'	230	na, nap, nai
bezweifeln	'to doubt'	301	na, ns-dass, ns-ob
bleiben	'to stay', 'to remain'	20,082	n, k
brechen	'to break'	786	n, na, nad, nar
denken	'to think'	3,293	n, na, np, ns-2
entziehen	'to take away'	410	nad, ndr
irren	'to be mistaken'	276	n, nr
klammern	'to cling to'	49	npr
lernen	'to learn'	1,820	n, na, ni
mangeln	'to lack'	438	x, xd, xp
scheinen	'to shine', 'to seem'	4,917	n, ni
stehlen	'to steal'	392	na, nad, nap
sträuben	'to resist'	86	nr, npr

Table 7: Lexical entries for verb subcategorisation

Lexicon Entry			
Verb		Freq	Subcategorisation
beauftragen	'to order', 'to charge'	230	na, nap:Dat.mit, nai
denken	'to think'	3,293	n, na, np:Akk.an, ns-2
enden	'to end'	1,900	n, np:Dat.mit
ernennen	'to appoint'	277	na, nap:Dat.zu
fahnden	'to search'	163	np:Dat.nach
klammern	'to cling to'	49	npr:Akk.an
schätzen	'to estimate'	1,357	na, nap:Akk.auf
stapeln	'to pile up'	137	nr, npr:Dat.auf, npr:Dat.in
sträuben	'to resist'	86	nr, npr:Akk.gegen
tarnen	'to camouflage'	32	na, nr, npr:Nom.vgl

Table 8: Lexical entries for verb subcategorisation including PP refi nement

7. Related Work

Automatic induction of subcategorisation lexica has mainly been performed for English. (Brent, 1993) used unlabelled corpus data and defi ned morpho-syntactic cues followed by a statistical fi ltering, to obtain a verb lexicon with six different frame types, without prepositional phrase refi nement. Brent evaluated the learned subcategorisation frames against hand judgements and achieved an f-score of 73.85%. (Manning, 1993) also worked on unlabelled corpus data and did not restrict the frame defi nitions. He applied a stochastic part-of-speech tagger, a fi nite state parser, and a statistical fi ltering process (following Brent). Evaluating 40 randomly selected verbs (out of 3,104) against *The Oxford Advanced Learner's Dictionary* (Hornby, 1985) resulted in an f-score of 58.20%. (Briscoe and Carroll, 1997) pre-defi ned 160 frame types (including prepositional phrase defi nitions). They applied a tagger, lemmatiser and parser to unlabelled corpus data; from the parsed corpus they extracted subcategorisation patterns, classifi ed and evaluated them, in order to build the lexicon. The lexical defi nitions were evaluated against the Alvey NL Tools dictionary (Boguraev et al., 1987) and the COMLEX Syntax dictionary (Grishman et al., 1994) and achieved an f-score of 46.09%. The work in (Carroll and Rooth, 1998) is closest to ours, since they utilised the same statistical grammar framework for the induction of subcategorisation frames, not including prepositional phrase defi nitions. Their evaluation for 200 randomly chosen verbs with a frequency greater than 500 against *The Oxford Advanced Learner's Dictionary* obtained an f-score of 76.95%.

For German, (Eckle, 1999) performed a semi-automatic acquisition of subcategorisation information for 6,305 verbs. She worked on annotated corpus data and defi ned linguistic heuristics in the form of regular expression queries over the usage of 244 frame types including PP defi nitions. The extracted subcategorisation patterns were judged manually. Eckle performed an evaluation on 15 hand-chosen verbs; she does not cite explicit recall and precision values, except for a subset of subcategorisation frames. (Wauschkuhn, 1999) constructed a valency dictionary for 1,044 verbs with corpus frequency larger than 40. He extracted a maximum of 2,000 example sentences for each verb from annotated corpus data, and constructed a context-free grammar for partial parsing. The syntactic analyses provided valency patterns, which were grouped in order to extract the most frequent pattern combinations. The common part of the combinations defi ned a distribution over 42 subcategorisation frame types for each verb. The

evaluation of the lexicon was performed by hand judgement on seven verbs chosen from the corpus. Wauschkuhn achieved an f-score of 61.86%.

Comparing our subcategorisation induction with existing approaches for English, (Brent, 1993; Manning, 1993; Carroll and Rooth, 1998) are more flexible than ours, since they do not require a pre-definition of frame types. But none of them includes the definition of prepositional phrases, which makes our approach the more fine-grained version. (Brent, 1993) outperformed our approach by an f-score of 73.85%, but only on six different frame types; (Manning, 1993) and (Briscoe and Carroll, 1997) both have f-scores below ours, even though the evaluations were performed on more restricted data. (Carroll and Rooth, 1998) reached the best f-score of 76.95% compared to 72.05% in our approach, but their evaluation was facilitated by restricting the frequency of the evaluated verbs to more than 500.

Compared to subcategorisation lexica for German, we do neither need extensive annotation of corpora, nor restrict the frequencies of verbs in the lexicon. In addition, our approach is fully automatic after grammar definition and does not involve massive heuristics or manual corrections. Finally, the evaluation was not performed by hand judgement, but rather extensively on independent manual dictionary entries.

8. Summary

We presented a large-scale computational subcategorisation lexicon for several thousand German verbs, unrestricted concerning the verb frequencies. The lexical entries were induced from a lexicalised probabilistic context-free grammar: we performed unsupervised training on the German grammar and developed a simple methodology to utilise frequency distributions in the resulting statistical grammar model as source for verb subcategorisation.

As to our knowledge, no former computational approach has obtained a subcategorisation lexicon for German comparable in size (the number of verbs in the lexicon), comprehension (no restrictions concerning the frequencies of the verbs), or verified reliability (evaluation on 3,090 verbs, without hand judgement). In addition, the type definition is variable with respect to the inclusion of prepositional phrase refinement.

The subcategorisation lexicon has been evaluated against dictionary definitions and proven reliable: the lexical entries hold a potential for adding to and improving manual verb definitions. The evaluation results justify the utilisation of the machine-readable lexicon as a valuable component for supporting NLP-tasks.

9. References

Leonard E. Baum. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities*, III:1–8.

B. Boguraev, E. Briscoe, J. Carroll, D. Carter, and C. Grover. 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.

Michael R. Brent. 1993. From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19:203–222.

Ted Briscoe and John Carroll. 1997. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*.

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

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

Judith Eckle. 1999. *Linguistic Knowledge for Automatic Lexicon Acquisition from German Text Corpora*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

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

A.S. Hornby. 1985. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, 4th edition.

K. Lari and S. J. Young. 1990. The Estimation of Stochastic Context-Free Grammars using the Inside-Outside Algorithm. *Computer Speech and Language*, 4:35–56.

Christopher D. Manning. 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.

Helmut Schmid. 2000. Lopar: Design and Implementation. Arbeitspapiere des Sonderforschungsbereichs 340 *Linguistic Theory and the Foundations of Computational Linguistics* 149, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, July.

Sabine Schulte im Walde, Helmut Schmid, Mats Rooth, Stefan Riezler, and Detlef Prescher. 2001. Statistical Grammar Models and Lexicon Acquisition. In Christian Rohrer, Antje Rossdeutscher, and Hans Kamp, editors, *Linguistic Form and its Computation*. CSLI Publications, Stanford, CA.

Sabine Schulte im Walde. 2002. Evaluating Verb Subcategorisation Frames learned by a German Statistical Grammar against Manual Definitions in the *Duden* Dictionary. In *Proceedings of the 10th EURALEX International Congress*, Copenhagen, Denmark. To appear.

Oliver Wauschkuhn. 1999. *Automatische Extraktion von Verbvalenzen aus deutschen Textkorpora*. Ph.D. thesis, Institut für Informatik, Universität Stuttgart.