Acquiring German Prepositional Subcategorization Frames from Corpora

Erika F. de Lima
GMD - German National Research Center
for Information Technology
Dolivostrasse 15
64293 Darmstadt, Germany
delima@darmstadt.gmd.de

July 7, 1997

Abstract

This paper presents a procedure to automatically learn German prepositional subcategorization frames from text corpora. It is based on shallow parsing techniques employed to identify high-accuracy cues for prepositional frames, the EM algorithm to solve the PP attachment problem implicit in the task, and a method to rank the evidence for subcategorization provided by the collected data.

1 Introduction

The description of lexical forms in both computation and human-oriented lexica include *prepositional subcategorization* information. For instance in German, the verb *arbeiten* ('to work') subcategorizes for a PP headed by the preposition *an* ('on'), and the verb *erinnern* ('to remind'), for an accusative NP and a PP headed by *an*:

(1) Mary arbeitet an der Frage $P \stackrel{?}{=} NP$.
Mary works on the question

(2) Mary erinnert ihren Freund an den Termin.
Mary reminds her friend on the deadline
'Mary reminds her friend of the deadline.'

Subcategorization information is usually compiled by hand. A procedure to automatically learn prepositional subcategorization would enable the acqui-

sition of broad-coverage lexica which reflect evolving usage and which are less subject to lexical gaps.

Learning prepositional subcategorization automatically is not a trivial task; it entails a PP attachment decision problem, and requires being able to distinguish complement from adjunct prepositional cues. For instance in (2) above, it is (syntactically) possible to attach the prepositional phrase [$_{PP}$ an den Termin] (to the noun phrase object as well as to the verb phrase. Sentence (2) cannot be considered conclusive evidence of a verbal frame based on syntactical information alone.

In (3) the prepositional phrase [$_{PP}$ in der Nacht] ('at night') is an adjunct PP which may occur with any (aspectually compatible) verb. It is not specific of the verb *arbeiten* ('to work') and should not be considered evidence of subcategorization.

(3) Mary arbeitete in der Nacht.
Mary worked in the night
'Mary worked at night.'

This paper proposes a method to automatically acquire German prepositional subcategorization frames (SFs) from text corpora. It is based on shallow parsing techniques employed to identify high-accuracy cues for prepositional SFs, and a method to rank the evidence for subcategorization provided by the collected data. The PP attachment problem implicit in the task is dealt with by using the EM algorithm to rank alternative frames. The subcategorization frames considered are shown in figure 1.

2 Method

The automatic extraction of German prepositional SFs is based on the observation that certain constructs involving so-called *pronominal adverbs* are high-accuracy cues for prepositional subcategorization. Pronominal adverbs are compounds in German consisting of the adverbs *da(r)-* and *wo(r)-* and certain prepositions. For instance in (4c), the pronominal adverb *daran* ('about it') is used as a pro-form for the personal pronoun *es* ('it') as the object of the preposition *an* ('about'). (Note that the usage of the pronoun (4b) is ungrammatical.) In (4d), the pronominal adverb *daran* occurs in a correlative construct with a subordinate *daß* ('that') clause immediately following it.

(4) a. Mary denkt an Johns Ankuft.
Mary thinks on John's arrival
'Mary thinks about John's arrival .'

Example	SF Description
PP[auf] V[warten] 'wait for'	Verb with PP
PP[an] NP_A V[erinnern] 'remind NP of'	Verb with accusative object and PP
PP[für] NP_D V[danken] 'thank NP for'	Verb with dative object and PP
PP[auf] sich V[vorbereiten] 'to prepare oneself for'	reflexive verb with PP
PP[auf] N[Hoffnung] 'hope for'	Noun with PP
PP[auf] A[stolz] 'proud of'	Adjective with PP

Figure 1: Subcategorization frames learned by the system

b. *Mary denkt an es.
Mary thinks on it
c. Mary denkt daran.
Mary thinks on it
'Mary thinks about it.'
d. Mary denkt daran, daß John bald ankommt.
Mary thinks on it that John soon arrives
'Mary thinks about the fact that John will arrive soon.'

Unlike prepositional phrases, pronominal adverb correlative constructs provide reliable cues for prepositional subcategorization. For instance the occurrence of the pronominal adverb *daran* in the correlative construct in (4d) can be used to infer that the verb *denken* ('to think') subcategorizes for a PP headed by the preposition *an* ('about').

In the next section, a learning procedure is described which makes use of pronominal adverb correlative constructs to infer prepositional subcategorization. It consists of four components: SF detection, mapping, disambiguation, and ranking.

2.1 SF Detection

This component makes use of shallow parsing techniques to detect possible prepositional SF structures; a standard CFG parser is used with a hand-written grammar defining pairs of main and subordinate clauses in correlative constructs such as (4d). Main clauses covered by the grammar include copular constructs as well as active and passive verb-second and verb-final constructs. Subordinate clauses considered include those headed by *daß* ('that'), indirect interrogative clauses, and infinitival clauses.

The internal structure of the clause pair consists of phrase-like constituents; these include nominative (NC), prepositional (PC), adjectival (AC), verbal (VC), and clausal constituents. Their definition is non-standard; for instance, all prepositional phrases, whether complement or not, are left unattached. As an example, the shallow parse structure for the sentence fragment in (5) is shown in (5′) below.

(5) Er lobte die Reaktion der öffentlichen Meinung in Rußland
he praised the reaction the public opinion in Russia
als Beweis dafür, daß...
as proof for it that
'He praised the reaction of the public opinion in Russia as proof of the fact that ...'

(5′) [$_S$ [$_{NC}$ Er]
[$_{VC}$ lobte]
[$_{NC}$ die Reaktion]
[$_{NC}$ der öffentlichen Meinung]
[$_{PC}$ in Rußland]
[$_{PC}$ als Beweis]
[$_{PC}$ dafür]
[$_{SC}$ daß...]
]

2.2 SF Mapping

The SF Mapping component maps a shallow parse structure of a main clause in a pronominal adverb correlative construct to a set of putative subcategorization frames reflecting structural as well as morphological ambiguities in the original sentence. Alternative SFs usually stem from an ambiguity in the attachment of the pronominal adverb PP. The mapping is defined as follows. (In the following, p denotes the preposition within the pronominal adverb

in a correlative construct main clause, VC the main verbal constituent in the clause; v in VC[v] denotes the head lemma of the verbal constituent, analogously for NC[n].)

VC[v]/NC[n]. An active verb-second or verb-final clause with one NC is mapped to {PP[p] V[v]} if the NC precedes the finite verb/auxiliary in the clause, otherwise to {PP[p] V[v], PP[p] N[n]}.

For instance, sentence (6) is a verb-second clause with an adverbial in the first position in the clause and one NC following the verb. In this construct, the PP headed by the pronominal adverb may potentially be attached to the verb phrase or to the nominal phrase immediately preceding it. According to this rule, this sentence is mapped to {PP[an] V[arbeiten], PP[an] N[Student]}.

(6) Jetzt arbeitet der Student daran, ...
Now works the student on it
'The student is now working on ...'

VC[v]/NC$_1$[n_1]/NC$_2$[n_2]. An active verb-second or verb-final clause with two nominal constituents NC$_1$ and NC$_2$ such that NC$_2$ follows NC$_1$ in the clause is mapped to {PP[p] NP$_A$ V[v], PP[p] N[n_2]}, if the head of NC$_2$ is a noun, and to {PP[p] NP$_A$ V[v]} otherwise.

Sentences (7a,b) are examples to which this rule applies. In (7a) the verb *erinnern* ('to remind') subcategorizes for an accusative NP and a PP headed by the preposition *an* ('on'), while in (7b), the verb *nehmen* ('to take') is a support verb and *Rücksicht* ('consideration') a noun which subcategorizes for a PP headed by the preposition *auf*. Since their shallow structure is ambiguous, they are each mapped to a SF set reflecting both attachment alternatives; (7a) is mapped to the set {PP[an] NP$_A$ V[erinnern], PP[an] N[Freund]}, and (7b) to the set {PP[auf] NP$_A$ V[nehmen], PP[auf] N[Rücksicht]}.

(7) a. Mary erinnert ihren Freund daran, daß...
Mary reminds her friend on it that
'Mary reminds her friend of the fact that ...'
b. Mary nimmt keine Rücksicht darauf, daß...
Mary takes no consideration on it that
'Mary shows no consideration for the fact that ...'

Copula/NC$_1$[n_1]/NC$_2$[n_2]. A copula clause with two nominal constituents NC$_1$[n_1] and NC$_2$[n_2] such that NC$_2$ follows NC$_1$ and n_2 is a noun is mapped to {PP[p] N[n_2]}. For instance (8) is mapped with this rule to {PP[auf] N[Hinweis]}.

(8) Weil dies ein Hinweis darauf ist, daß...
because this an indication on in is that
'Because this is an indication (of the fact) that ...'

Copula/NC[n]/AC[a]. A copula clause with one nominal and one adjectival constituent is mapped to {PP[p] N[n], PP[p] A[a],}. For instance, with this rule the clause in (9) is mapped to {PP[auf] A[stolz], PP[auf] N[Student]}

(9) Stolz ist der Student darauf, daß...
proud is the student on it that
'The student is proud of the fact that ...'

PCs. Any clause in which a PC immediately precedes the pronominal adverb is mapped as in the appropriate rule with the additional element 'PP[p] N[n]' in the set, where n is the head of the NC within the prepositional constituent. For instance, (10) is mapped to {PP[an] V[arbeiten], PP[an] N[Woche]} with the VC/NC and PC rules.

(10) Mary arbeitet seit zwei Wochen daran, ...
Mary works since two weeks on it
'Mary has been working for two weeks on ...'

Morphology. Any clause in which a possible locus of attachment is morphologically ambiguous is mapped with the appropriate rule applied to all morphology alternatives. For instance, (11) is mapped with the VC/NC and Morphology rules to {PP[an] V[denken], PP[an] V[gedenken]}, since *gedacht* is the past participle of both the verbs *denken* ('to think') and *gedenken* ('to consider').

(11) Er hat daran gedacht, daß ...
he has on it thought/considered that
'He thought of ...'

Passive/VC[v]/NC[n]. This rule is applied to *werden* ('to be') passive verb-second or verb-final clause with one NC. In case n is not the pronoun *es* ('it'), the clause is mapped to {PP[p] NP_A V[v]} if NC precedes the verb, and to {PP[p] NP_A V[v], PP[p] N[n]} otherwise. In case n is the pronoun *es*, the clause is mapped to {PP[p] NP_A V[v], PP[p] V[v]}. For instance, (12) is mapped to {PP[an] NP_A V[erinnern]}.

(12) Mary wird daran erinnert, daß...
Mary is on it reminded that
'Mary is reminded (of the fact) that ...'

2.3 SF Disambiguation

The disambiguation component uses the expectation-maximization (EM) algorithm to assign probabilities to each frame in an SF alternative, given all SF sets obtained for a given corpus. The EM algorithm (Dempster, Laird, and Rubin, 1977) is a general iterative method to obtain maximum likelihood estimators in incomplete data situations. See (Vardi and Lee, 1993) for a general description of the algorithm, as well as numerous examples of its application. The EM algorithm has been used to induce valence information in (Carrol and Rooth, 1997).

In the current setting, the algorithm is employed to rank the frames in a given SF set by using the relative evidence obtained for each frame in the set. The algorithm is shown below.

Algorithm. Let F be a set of *frames*. Further, let $\mathcal{S}$ be a finite set of nonempty subsets of $\wp(F)$, and let $F_0 = \bigcup_{X \in \mathcal{S}} X$.

Initialization step: for each frame x in F_0:

$$c_0(x) = \sum_{X \in \mathcal{S}} (I(x, X) \cdot g_C(X))$$

Step $k+1$ ($k >= 0$):

$$c_{k+1}(x) = c_k(x) + \sum_{X \in \mathcal{S}} (P_k(x, X) \cdot g_C(X))$$

Where g_C is a function from $\mathcal{S}$ to the natural numbers mapping a set X to the number of times it was produced by the SF mapping for a given corpus C. Further, I, P_k, and p_k are functions defined as follows:

$$I: \; F \times \wp(F) \rightarrow [0,1]$$

$$(x, X) \mapsto \begin{cases} \frac{1}{|X|} & \text{if } x \in X \\ 0 & \text{else} \end{cases}$$

$$P_k: \; F \times \wp(F) \rightarrow [0,1]$$

$$(x, X) \mapsto \begin{cases} \frac{p_k(x)}{\sum_{\bar{x} \in X} p_k(\bar{x})} & \text{if } x \in X \text{ and } |X| > 1 \\ 0 & \text{else} \end{cases}$$

$$p_k: \; F \rightarrow [0,1]$$

$$x \mapsto \frac{c_k(x)}{\sum_{\bar{x} \in F_0} c_k(\bar{x})}$$

Definition. A frame x is *best in the set X at the iteration k* if $x \in X$ and $p_k(x)$ is an absolute maximum in $\bigcup_{\bar{x} \in X} p_k(\bar{x})$.

In the algorithm above, $\mathcal{S}$ denotes the set of SF sets produced by the SF mapping for a given corpus C. In the initialization step, c_0 assigns an initial "weight" to each frame, depending on its relative frequency of occurrence, and on whether the structures in which it occurred are ambiguous. The weight $c_k(x)$ of a frame x is used to estimate its probability $p_k(x)$. In each iteration of the algorithm, the weight of a frame x is calculated by considering the totality of alternatives in which x occurs (i.e., the sets for which $x \in X$ and $|X| > 1$), and its probability within each alternative. The best frames in a set are the most probable frames given the evidence provided by the data. In the experiment described in section 3, the final number of iterations was set empirically.

2.4 SF Ranking

This component ranks the SFs obtained by the previous component of the system. Let $\mathcal{L}_C$ be the set of head lemmata (verbs, nouns and adjectives) in the subcategorization cues (i.e., best frames in the SF sets) for a given corpus C. Let $\mathcal{F}$ be the set {NP_A V[·], NP_D V[·], V[·], PP[an] V[·], PP[an] NP_A V[·], ...} of *SF structures*. (Roughly, an SF structure is an SF without its head lemma.) The analysis of SF cues is performed by creating a contingency table containing the following counts for each lemma $L \in \mathcal{L}_C$ and prepositional structure $S \in \mathcal{F}$: $k(L\,S)$ ($k(L\,\bar{S})$) is the count of lemma L with (without) structure S, and $k(\bar{L}\,S)$ ($k(\bar{L}\,\bar{S})$) is the count of all lemmata in $\mathcal{L}_C$ except L with (without) structure S.

If a lemma L occurs independently of a structure S, then one would expect that the distribution of L given that S is present and that of L given that S is not present have the same underlying parameter. The log likelihood statistic is used to test this hypothesis. This statistic is given by $-2\log\lambda = 2(\log L(p_1,k_1,n_1)+\log L(p_2,k_2,n_2)-\log L(p,k_1,n_1)-\log L(p,k_2,n_2))$, where $\log L(p,k,n) = k\log p + (n-k)\log(1-p)$, and $p_1 = \frac{k_1}{n_1}$, $p_2 = \frac{k_2}{n_2}$, $p = \frac{k_1+k_2}{n_1+n_2}$ (For a detailed description of the statistic used, see (Dunning, 1993)).

In the formulae above, k_1 is $k(L\,S)$, n_1 is the total number of occurrences of S, k_2 is $k(L\,\bar{S})$, and n_2 the total number of occurrences of structures other than S. A large value of $-2\log\lambda$ for a lemma L and structure S means that the outcome is such that the hypothesis that the two distributions have the same underlying parameter is unlikely, and that a lemma L is highly associated with a structure S in a given corpus. This value is used to rank the subcategorization cues produced by the previous components of the system.

3 Results

The method described in the previous section was applied to 1 year of the newspaper *Frankfurter Allgemeine Zeitung* containing approximately 36 million word-like tokens. A total of 16795 sentences matched the pronominal adverb correlative construct grammar described in section 2.1.

3.1 SF Disambiguation

Of the 16795 sets produced by the SF mapping, 5581 contained more than one SF, i.e., reflected some form of ambiguity in the original sentence, of which 4365 were unique. A random set of 400 sets was obtained from these unique ambiguous sets. The disambiguation component produced a decision for 359 of these 400 sets. These results were compared to the blind judgements of a single judge; 305 were found to be correct, 23 incorrect. The remaining 31 sets were considered to contain incorrect SFs solely. Although an error rate of over 15% is not negligible, it is comparable to other PP attachment experiments (Collins and Brooks, 1995).

3.2 Acquired Dictionary

The system acquired a dictionary of 1663 unique subcategorization frames. Figure 2 and 3 show the 30 most and 10 least plausible frames according to the system. Starred structures are considered to be errors.

Examination of the ranked SF table shows that frames with a low $-2 \log \lambda$ value consist mostly of errors. The cues produced by the system are not perfect predictors of subcategorization. False cues stem from incorrect decisions in the disambiguation component as well as parsing and mapping errors, spurious adjuncts, or actual errors in the original text.

In figures 3, two errors are due to the disambiguation component (*nehmen, Amt*); three errors stem from mistaking reflexive verbs for verbs taking any accusative object (*sich treffen mit* ('to meet with'), *sich bekennen zu* ('declare oneself for'), *sich halten an* ('to comply with')). These stem from the grammar specification, and can be avoided with further development of the detection component.

By far the most frequent type of error was the inclusion of an accusative or dative NP in a verbal frame when the verb in fact only takes a PP. For instance of the errors in the 31 sets (out of the 400 ambiguous sets examined) containing incorrect SFs only, about 42% were due to the fact that an additional accusative/dative NP was incorrectly included in a verbal frame,

$-2\log\lambda$	$k(L\,S)$	$k(\bar{L}\,S)$	$k(L\,\bar{S})$	$k(\bar{L}\,\bar{S})$	L	S
13270.3167	1225	1691	2842	3897009	hinweisen	PP[auf] V[·] ('point to')
6757.6162	498	337	1183	3900749	ausgehen	PP[von] V[·] ('assume')
4857.2234	482	328	7909	3894048	rechnen	PP[mit] V[·] ('reckon with')
4241.0161	429	529	5792	3896017	erinnern	PP[an] V[·] ('remind of')
3307.6279	406	2510	3479	3896372	verweisen	PP[auf] V[·] ('refer to')
3179.3391	339	234	11293	3890901	bestehen	PP[in] V[·] ('lie in')
3156.5375	342	433	8013	3893979	sorgen	PP[für] V[·] ('care for')
3118.6878	255	158	2810	3899544	aussprechen	PP[für] sich V[·] ('speak for')
2897.6673	293	888	2766	3898820	beitragen	PP[zu] V[·] ('contribute to')
2548.1622	385	796	20598	3880988	führen	PP[zu] V[·] ('lead to')
2253.9826	234	2682	860	3898991	ankommen	PP[auf] V[·] ('depend on')
2002.4658	174	128	3706	3898759	begründen	PP[mit] NP_A V[·] ('substantiate NP with')
1605.2355	146	629	1155	3900837	plädieren	PP[für] V[·] ('plead for')
1193.6521	190	383	25066	3877128	liegen	PP[in] V[·] ('lie in')
1042.8259	115	298	4968	3897386	einsetzen	PP[für] sich V[·] ('support')
876.6903	64	121	610	3901972	hindern	PP[an] NP_A V[·] ('hinder NP from')
813.5798	74	761	510	3901422	abhängen	PP[von] V[·] ('depend on')
789.1838	78	122	4432	3898135	Hinweis	PP[auf] N[·] ('reference to')
777.0291	121	837	7860	3893949	denken	PP[an] V[·] ('think of')
776.2428	62	92	1368	3901245	aufmerksam	PP[auf] A[·] ('attentive to')
766.8966	122	1059	6794	3894792	dienen	PP[zu] V[·] ('serve for')
766.6407	65	89	2105	3900508	stolz	PP[auf] A[·] ('proud of')
764.9588	135	640	16398	3885594	sprechen	PP[für] V[·] ('speak for')
686.0675	48	393	107	3902219	hinweg-täuschen	PP[über] V[·] ('obscure')
684.1054	85	40	27806	3874836	sehen	PP[in] NP_A V[·] ('see NP in')
677.7402	70	1111	660	3900926	neigen	PP[zu] V[·] ('tend to')
656.8564	67	455	1435	3900810	Beweis	PP[für] N[·] ('proof of')
577.1696	58	383	1371	3900955	nachdenken	PP[über] V[·] ('think about')
569.0955	43	78	815	3901831	abhalten	PP[von] NP_A V[·] ('prevent NP from')
555.6635	61	89	7787	3894830	Interesse	PP[an] N[·] ('interest in')

Figure 2: 30 most plausible frames

$-2\log\lambda$	$k(L\,S)$	$k(\bar{L}\,S)$	$k(L\,\bar{S})$	$k(\bar{L}\,\bar{S})$	L	S
0.0126	1	301	11527	3890938	treffen	*PP[mit] NP_A V[·]
0.0117	1	463	9351	3892952	bekennen	*PP[zu] NP_A V[·]
0.0112	4	831	17723	3884209	wissen	PP[von] V[·] ('know of')
0.0087	1	204	20866	3881696	nehmen	*PP[auf] NP_A V[·]
0.0054	1	159	22650	3879957	finden	*PP[durch] NP_A V·[]
0.0047	1	184	22565	3880017	halten	*PP[an] NP_A V[·]
0.0029	1	809	5082	3896875	einsetzen	*PP[mit] V[·]
0.0011	1	957	3938	3897871	verdienen	PP[an] V[·] ('make a profit on')
0.0005	1	204	18633	3883929	bringen	PP[auf] NP_A V[·]
0.0002	1	521	7580	3894665	Amt	*PP[für] N[·]

Figure 3: 10 least plausible frames

although the preposition in the frame was subcategorized for. These stem from erroneous alternatives in the segmentation of nominal constituents as defined by the grammar and could be eliminated with further developed of the detection component.

Yet another type of error stems from pronominal adverbs which are conjunction/adverb homographs, or which are used anaphorically, while the verb in the main clause subcategorizes for a *daß* ('that') clause, so the sentence is erroneously considered to be a correlative construct. This is the source of most errors for frames involving the preposition *gegen* ('against'), *bei* ('by') and *nach* ('to'), and cannot be avoided given the learning strategy.

Given the fact that the cues produced by the system are not perfect predictors of subcategorization, a test of significance could be introduced in order to filter out potentially erroneous cues. However, it was observed that truly "new" prepositional frames—frames not listed in broad coverage published dictionaries, or even considered to be erroneous by a native speaker until confronted with examples from the corpus—behaved with respect to their rankings very much like errors. So the current version of the learning procedure relies on manual post-editing assisted by the SF ranking and examples from the corpus in order to discard false frames.

3.3 Precision and Recall

Evaluating the acquired dictionary is not straightfoward; linguists often disagree on the criteria for the complement/adjunct distinction. Instead of

attempting a definition, the acquired dictionary was compared to a broad coverage published dictionary containing explicit information on prepositional subcategorization.

A random set of 300 verbs occurring more than 1000 times in the corpus was obtained.[1] The prepositional SFs for these verbs which were listed in (Wahrig, Kraemer, and Zimmerman, 1980) and in the acquired lexicon were noted. There was a total of 307 verbal prepositional frames listed in either dictionary. Of these, 136 were listed only in the published dictionary, and 121 only in the acquired dictionary.

These prepositional SFs were used to calculate a lower bound for the precision and recall rates of the system; A SF is considered *correct* if and only if it is listed in the published dictionary.[2] A lower bound for the recall rate of the system is given by the number of learned correct frames divided by the number of frames listed in the published dictionary, or 52/173. This recall rate is a lower-bound for the actual rate with respect to the corpus, since there are prepositional SFs listed in the published dictionary with no instance in the corpus.

A lower bound for the precision of the system is given by the number of learned correct frames divided by the number of learned frames, or 52/188. This rate is a lower-bound for the actual precision rate of the system, since it does not take the fact into account that the system did learn true SFs not listed in the published dictionary, so the precision rate of the system is actually higher. Further, not all prepositions contributed equally to the precision and recall rates. For instance the precision and recall for the prepositions *aus* ('out') was 60% and 42%, that of *von* ('of') 50% and 53%, while that of *gegen* ('against') 6% and 11%, respectively.

4 Related Work

The automatic extraction of English subcategorization frames has been considered in (Brent, 1991; Brent, 1993), where a procedure is presented that takes untagged text as input and generates a list of verbal subcategorization frames. The procedure uses a very simple heuristics to identify verbs; the syntactic types of nearby phrases are identified by relying on local morpho-syntactic cues. Once potential verbs and SFs are identified, a final com-

[1] There was a total of 15178 unique verbs (known to the morphology) occurring in the corpus, of which 913 occurred more than 1000 times.

[2] No dictionary is exempt from errors (of omission). However it (hopefully) provides a uniform classification for PP subcategorization.

ponent attempts to determine when a lexical form occurs with a cue often enough so that it is unlikely to be due to errors; an automatically computed error rate is used to filter out potentially erroneous cues. Prepositional frames are not considered, since, according to the author, "it is not clear how a machine learning system could do this [determine which PPs are arguments and which are adjuncts]."

In (Manning, 1991) another method is introduced for producing a dictionary of English verbal subcategorization frames. This method makes use of a stochastic tagger to determine part of speech, and a finite state parser which runs on the output of the tagger, identifying auxiliary sequences, noting putative complements after verbs and collecting histogram-type frequencies of possible SFs. The final component assesses the frames encountered by the parser by using the same model as (Brent, 1993), with the error rate set empirically. Prepositional verbal frames are learned by the system by relying on PPs as cues for subcategorization; since the system cannot differentiate between complement and adjunct prepositional cues, it learns frequent prepositional adjuncts as well.

In order to evaluate the acquired dictionary, Manning compares the frames obtained for 40 random verbs to those in a published dictionary, yielding for these verbs an overall precision and recall rates of 90% and 43% respectively. However, if only the prepositional frames listed for these verbs are considered, the rates drop to approximately 84% and 25%, respectively. In the experiment described, the error bounds for the filtering procedure were chosen with the aim of "get[ing] a highly accurate dictionary at the expense of recall." His system did not consider nominal and adjectival frames.

(Carrol and Rooth, 1997) present a learning procedure for English subcategorization information. Unlike previous approaches, it is based on a probabilistic context free grammar. The system uses expected frequencies of head words and frames—calculated using a hand-written grammar and occurrences in a text corpus—to iteratively estimate probability parameters for a PCFG using the expectation maximization algorithm. These parameters are used to characterize verbal, nominal and adjectival SFs. The model does not distinguish between complements and adjunct prepositional cues.

5 Conclusion

This paper presents a method for learning German prepositional subcategorization frames. Although other attempts have been made to learn English verbal/prepositional SFs from text corpora, no previous work considered a

partially free word-order language such as German, nor differentiated between complement and adjunct prepositional cues.

The overall precision rate for the system described in this paper is lower than that of similar systems developed for English, since no test of significance was used to filter out possibly erroneous cues. In the experiment described in the previous section, truly new prepositional frames behaved with respect to frequency of occurrence very much like errors, and would possibly have been discarded by a filtering mechanism.

A problem in the current version of the system was the fact that segmentation of nominal constituents was not optimally handled by the detection component, leading to a large number of verbal frames with correct prepositions, but with an additional erroneous accusative/dative NC in the frame. So the precision of the system can be significantly improved with further development of the detection component.

Further, the system should be extended to handle other types of pronominal adverb cues, such as pro-forms for interrogative, personal and relative pronouns; possibly PPs headed by prepositions should also be considered. Finally, the method–low-level parsing together with a procedure to rank alternatives obtained–should be extended to other frames as well.

References

Brent, Michael R. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeding of the ACL*, pages 209–214.

Brent, Michael R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262.

Carrol, Glenn and Mats Rooth. 1997. Valence induction with a head-lexicalized PCFG. http://www2.ims.uni-stuttgart.de/~mats.

Collins, Michael and James Brooks. 1995. Prepositional phrase attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora.*

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from inclomplete data via the em algorithm. *J.R.Statis. Soc. B*, 39:1–38.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Manning, Christopher D. 1991. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 29th Annual Meeding of the ACL*, pages 235–242.

Vardi, Y. and D. Lee. 1993. From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems. *J.R.Statis. Soc. B*, 55(3):569–612.

Wahrig, Gerhard, Hildegard Kraemer, and Harald Zimmerman. 1980. *Brockhaus Wahrig Deutsches Wörterbuch in sechs Bänden.* F.A. Brockhaus und Deutsche Verlags-Anstalt GmbH, Wiesbaden.