Extracting raw material for a German subcategorization lexicon from newspaper text

Judith Eckle, Ulrich Heid

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart, Germany

Abstract

This paper is about extracting evidence for syntactic subcategorization phenomena from German newspaper text. The purpose of this work is to support and partly automatize the construction of a subcategorization lexicon for NLP, similar, for example, to COMLEX.
We here report on the extraction of verb lists and sample sentences illustrating syntactic construction possibilities. The lists are ordered by subcategorization types; they are manually screened to remove noise, and then used to automatically produce proto-entries of the lexicon.
Since no phrasal parsing is yet available for German, we use part-of-speech shapes (a regular grammar over categorially and morphosyntactically annotated word forms) and lemma information; to reduce the noise produced by general part-of-speech shapes, we have defined "constraining contexts" and use a context-dependent modeling.

The retrieval results contain less than 5% of noise. Moreover, we can retrieve specific types of syntactic information which cannot be found in any traditional dictionary: we can, for example, identify verbs with "obligatory coherent" infinitives (cf. [Haider 1993]).

We explain the principles and procedures of our extraction work, discuss the case of infinitive-taking verbs and assess the results obtained on the first 3.000 readings extracted.

1 Introduction

1.1 Motivation and Approach

There is no freely available electronic subcategorization dictionary of German yet[1]. A typical example of this kind of dictionary is COMLEX, for English. It contains detailed syntactic information in a format which supports in principle the reformatting towards different specific representations, as they are used in computational linguistic grammar formalisms (cf. [Grishman/MacLeod/Meyers 1994], [Grishman/MacLeod 1994]). It has mostly been created manually, from machine-readable dictionaries and NLP lexicons.

The goal of the work we report on here is, however, to provide as much raw material for a subcategorization dictionary of German verbs (and later adjectives and nouns) as possible by automatic means. The dictionary construction itself still necessitates human intervention, but a large part of the preparatory work is done automatically.

For such a task, one could make use of low-level parsing (phrase types); for English and French, robust broad coverage grammars for the identification of phrasal categories are available (for example the *Fiddich* parser, cf. [Hindle 1991], or the English Constraint Grammar, cf. [Voutilainen et al. 1992]). But no similar tools are yet available for German, and parsing results must (in part) be simulated through the use of part-of-speech shapes, i.e. sequences of categorially and morphosyntactically annotated word forms. The part-of-speech shapes must be specific, to avoid too much noise in the extraction result. Thus we search for "constraining contexts", i.e. sentences where certain phrases or sequences of annotated word forms can only and unambigously be interpreted as illustrating exactly one given subcategorization pattern of a verb.

To prepare candidate verb lists, a set of queries is applied to a given text corpus; the lexicographer may select a query corresponding to a given construction[2]. The intermediate output consists of lemma frequency lists and sample sentences, both sorted by syntactic constructions (see section 3.1). The frequency lists indicate the absolute number of contexts unambigously illustrating a given syntactic construction of a given verb.
The lexicographer checks the lemma lists (by assessing the sample sentences) and decides for which lemmata a dictionary entry should be created. The resulting candidate list is input to a program which constructs proto-enties for each syntactic reading (i.e. pair of lemma and subcategorization frame).

[1]The PAROLE project, a development project financed partly by the European Commission, DG XIII E, Luxemburg, under the Language Engineering programme, will create such lexicons (of some 10.000 entries) for the major European languages.

[2]By running all available queries, evidence for the whole fragment (see below, section 4.1) can be found. Similarly, subcategorization in specialized language could be analysed, if a large enough corpus of specialized texts were available.

1.2 Infrastructure

Requirements for retrieval. The extraction of linguistic and lexicographic evidence from text corpora typically leads to large quantities of material. As in Information Retrieval, one has to ensure acceptable precision (the retrieved material must be relevant for the task expressed in the query, and there should be little noise) and an acceptable recall (as much as possible of the relevant material contained in the corpus should indeed be found, and there should be little silence).

We opted for high precision at the price of lower recall, accepting the fact that we may exclude some relevant candidates. The reasons for this procedure are that the elimination of noise must be done manually, which is time-consuming and expensive, and that the silence produced contains a large portion of material just corroborating the facts obtained with the restrictive approach.

The information available in corpora. Raw text material usable for the acquisition of linguistic knowledge is available for many languages[3]. But the degree to which the texts can be automatically annotated differs considerably between languages: other than for English, there is no robust phrase-level parsing yet for German[4]. Thus, for our extraction work, we can only use categorial information, morphosyntactic tags[5] and lemmatization results, introduced into the texts by means of the appropriate tools.

To extract evidence for syntactic constructions, we must rely on part-of-speech shapes and on lemma information. The extraction routines basically encode a regular grammar.

Tool infrastructure. Our extraction scenario comprises two main phases. The first one is linguistic pre-processing and automatic annotation of texts (see above), the second one is corpus query.

We use the following corpus query tools[6]:

- CQP, a general corpus query processor, for complex queries with any number and combination (including negation) of annotated information types, such as word forms, part-of-speech tags, lemmas, as well as possibly sentence or phrase boundaries.

[3]We make use, among others, of the following German newspapers: *Stuttgarter Zeitung* (special contract), *die tageszeitung* (CD-ROM), *Frankfurter Rundschau* (from the ECI CD-ROM). The material adds up to around 200 million tokens.

[4]The SPARKLE project (an Linguistic Engineering project financed partly by the European Commission, DG XIII, Luxembourg) will produce a chunk parser for German (cf.[Rooth/Carroll 1996]) in the medium term.

[5]Provided by the STTS tagset, an EAGLES-conformant morphosyntactic annotation scheme with 54 different tags; see [Schiller/Teufel/Thielen 1995] on STTS; the tagset (**S**tuttgart-**T**übingen **T**ag **S**et) is trivially mappable onto the EAGLES specifications for the morphosyntactic description of German, as described in [Teufel 1995]

[6]The corpus tools have been implemented in Perl, C, C++ and to some extent in UNIX-tools. TheXkwic user interface is also based on C and C++ and integrated into an X/MOTIF environment. The corpus queries are written in the CQP corpus query language which uses the standard posix-egrep regular expression notation. For details see [Schulze 1996]

- A macro processor for the CQP query language allowing to execute the same query on elements from lists. Moreover, query expressions can be named, stored and reused.
- XKWIC, an X Windows/MOTIF-based graphical user interface for the CQP corpus query language (cf. [Christ 1994b]) which provides keyword in context concordances, and allows to automatically sort the extracted material according to user-defined context parameters; lists of absolute and relative frequency of search items can be compiled.

2 Principles and Method

2.1 Motivation

In figure 1[7], we give an example of a simplistic extraction scheme for transitive verb candidates, along with, in the three rightmost columns, examples of both expected results (col. 3) and noise (cols. 4 and 5).

	Fact (1)	Encoding (2)	Examples (3)	(4)	(5)
a	Subord.conj.	[pos = "KOUS"]	daß	daß	daß
b	Article	[pos = "ART"]	der	die	die
c	Noun	[pos = "NN"]	Hund	Regierung	Zahl
d	Article	[pos = "ART"]	das	der	der
e	Noun	[pos = "NN"	Rennen	Forderung	Neuzulassungen
f	Verb candid.	[pos = "VVFIN"]	verlor	zustimmte	sinkt
g	within a sent.	within s	.</s>	.</s>	.</s>

Figure 1: Search for verbs with two NPs by means of part-of-speech shapes only

The examples show that mere pos-shapes (as given in column (2)) are not apt to capture transitive verb constructions, because they are not constrained enough: (4) is an example of a verb with an indirect object (*der* in box (4d) is a dative) and (5) is an example illustrating an intransitive verb (*sinken*; the NP in (5d/e) is a genitive attribute to the subject NP (5b/c)).

The queries must be more constrained. This is achieved through two types of devices, namely the search for constraining contexts, and a context dependent modeling of phrasal constructs.

[7] In this figure and in the subsequent analogous ones, we display the sentences from top to bottom; we usually give a paraphrase of the phenomen searched (column 1), the encoding used (col. 2) and examples.

2.2 Constraining contexts

To avoid noise in the extraction results, the queries used should only match contexts which unambiguously illustrate a given subcategorization frame of a verb. This implies searching for noun phrases or components thereof which have unambiguous morphosyntactic case marks. Not all noun forms have clearly identifiable case endings (cf. *Frauen* in figure 2), but many pronouns and determiners do have such forms (example: *einigen* in figure 2).

The table 1 in appendix A contains more examples of pronouns and determiners. These alone show that there will be some silence in the query results. Mostly, we have to rely on sentences with NPs whose head nouns are masculine, because many feminine and neuter forms are ambiguous.

However, some ambiguities do not cause problems in the extraction; for example, the ambiguity between accusative and nominative is not a major problem in the extraction of transitive verb evidence (see section 3.1); similarly, when extracting two-place constructions with other complements than direct (accusative) objects, it is sufficient to describe this complement unambiguously.

2.3 Context-dependent modeling

A number of parameters have to be kept track of to achieve a significant coverage. These include morphosyntactic properties of the verb under analysis (separable verb prefix: *die Entscheidung hängt von ihm ab* (*abhängen*); reflexive verbs: *er sorgt sich um seine Familie* (*sich sorgen*)), and more crucially, syntactic variation, such as the three different models of word order in German (see below, section 3.2). This leads to slight differences in the searchable pos-shape models for, e.g., NPs, depending on the word order model in question: an independent encoding, as in a normal analysis grammar and its reuse in all possible contexts would be more modular in design, but would either lead to much more silence (if the most restricted definitions were used) or to much more noise and ambiguous corpus samples (if more liberal definitions were used).

The sentences in figure 2 provide some illustration of this problem. The noun form *Frauen* is ambiguous with respect to case: it can have any of the four cases. The ambiguity does not cause problems when a subject NP with an intransitive verb is considered (see sentence (1)), because the context forces a nominative interpretation. It does lead to noise, however, in the extraction of two-place verbs, as illustrated by the retrieval of sentences (2) and (3), which are examples of a direct and an indirect object, respectively. Since NPs without determiner can lead to ambiguities of the kind of (2) vs. (3) with one and the same query, the queries for verbs with direct and indirect objects have been modified to include an obligatory determiner; unambiguous examples of this type are found in (4) and (5).

No.	conj	subject	complement	verb	case of compl	ambig.
(1)	weil	Frauen		kommen	nominative	(Y)
(2)	weil	er	Frauen	sieht	accusative	Y
(3)	weil	er	Frauen	vertraut	dative	Y
(4)	weil	er	einige Frauen	sieht	accusative	N
(5)	weil	er	einigen Frauen	vertraut	dative	N

Figure 2: Interaction between constraining contexts and context dependent modeling

3 The semi-automatic construction of lexicon entries – Examples

3.1 A simple example: transitive verbs

Queries. To find evidence for two-place transitive verbs, sentences containing one nominative NP and one accusative NP (both identified by the appropriate determiners) are retrieved. To reduce the amount of silence, the order of the NPs is left open by additionally allowing determiners which are ambiguous between nominative and accusative. So constituent order variation is captured, but two place predicative verbs (two nominatives: *sein*, *werden*, *heißen*, *bleiben*) must be explicitely removed from the result set. Depending on the word order type, a few additional constructions can be allowed in the NPs without introducing ambiguity (e.g. postnominal PPs in a noun phrase in the "Vorfeld" of a verb-second sentence).

Raw material: frequency tables. Figure 6 in annex B contains an extract from a frequency list of candidate verbs taking a nominative NP and an accusative NP[8]. To get a full picture of the distribution of a subcategorization scheme across a corpus, the frequency tables obtained from the analysis of different contexts (e.g. verb-first, verb-second, verb-final) need to be merged.

[8]The frequency list refers to verb-second sentences in present tense or imperfect (without separable prefix), in 200 million words of German newspaper text.

The frequency tables are relevant for the lexicographer, because in many cases, low frequency items include some noise. An example are verbs with subcategorized prepositional objects: usually, the most frequent preposition candidates for a given verb tend to be the prepositions governing a prepositional object ("semantically empty" complement prepositions), but the low frequencies contain adjunct prepositions[9]. For example, the frequency table for verbs with *an*-objects has *liegen an*, *erinnern an*, *glauben an*, *denken an* at the high frequency end, which all have semantically empty subcategorized prepositional objects; low frequency items include as well adjuncts such as *zerfallen (an der Luft)*, but also examples of prepositional objects, such as *gewöhnen an*. In such cases, the lexicographer should consult the examples and decide on the inclusion in the dictionary.

Sample sentences. Examples of sample sentences are given in figure 7, in annex B, for verbs taking a nominative and an accusative NP[10]. Repetitions in the set of samples are automatically detected: instead of displaying large amounts of analogous keyword in context (= kwic) concordances, we use a simple "condensation tool" which calculates the number of identical kwic matches and displays only one of them along with a frequency count.

Proto-entries. Once the lexicographer has decided which lemmas are admitted to the lexicon, proto-entries for the identified subcategorization readings are automatically produced. An example, illustrating the use of *fordern* with a nominative and an accusative NP, is displayed in figure 8, in annex B. Each record consists of the verb lemma ("<verb>...</verb>"), a subcategorization pattern corresponding to the query executed, and a set of randomly chosen example sentences[11].

[9]The same problem comes up here as in any descriptive linguistic work; there are no corpus-reproducible facts from where to derive any clear argument ↔ adjunct distinction. As a rule of thumb, we assume that highly frequent combinations tend to have argument status.

[10]They have been taken from the subset of a 200 million word newspaper corpus which contains subclauses in present or past tense (verb-final word order) introduced by conjunctions.

[11]If the "condensation tool" has provided frequency counts for contexts, the most frequent (most typical?) ones are selected.

3.2 Case Study: Verbs taking infinitives with *zu*

The Problem. In German, verbs taking infinitives with *zu* can occur in verb-last sentences in two different constructions:

(1) *..., weil Hans das Buch zu lesen versucht*: no extraposition [*zuinf fin*]

(2) *..., weil Hans versucht, das Buch zu lesen*: extraposition [*fin zuinf*]

In (1), the zu-infinitive comes before the finite verb, whereas in (2) extraposition of the zu-infinitive has taken place. For a number of verbs however, extraposition of the zu-infinitive is not possible[12], for example the verb *scheinen*:

(3) *..., weil Hans das Buch zu lesen scheint*: no extraposition [*zuinf fin*]

(4) *..., ⋆weil Hans scheint, das Buch zu lesen*: extraposition [*fin zuinf*]

As the possibility to take one or the other construction, or both, seems to be a lexical property of the respective verb, an NLP lexicon has to provide information about the possible constructions for each verb taking zu-infinitives. Since in traditional dictionaries such information is not available, it is worth while to extract it from text corpora.

Extraction procedure. We look for verbs taking zu-infinitives which do not allow extraposition of the zu-infinitive, i.e. of the type [*zuinf fin*] as in (3); we expect that these do not occur in constructions of type [*fin zuinf*] (see (2) and (4)). Text corpora however do not provide negative evidence: when a certain construction does not occur in the corpus, it can not be concluded that it is not possible. What we can extract therefore from corpora are lists of candidates with the behaviour of the lexical class of obligatorily coherent verbs.

Our approach to identify verb candidates is to extract two sets of verbs together with their frequency distributions and to compare them: firstly, a set of verbs in verb-last sentences where the zu-infinitive comes before the finite verb (set *zuinf-fin* and frequency distribution *freq-zuinf-fin*, see (1) and (3)) and secondly, a set of verbs in verb-last-sentences with extraposed zu-infinitive (set *fin-zuinf* and frequency distribution *freq-fin-zuinf*, see (2)).

Then the verbs we are looking for should occur only in set *zuinf-fin* and not in set *fin-zuinf*. A very simple method to get these verbs would be to compute the set of verbs which are only in *zuinf-fin* and not in *fin-zuinf*. But this does not take into account that the sets of extracted verbs could contain noise resulting for instance from tagging errors. So this simple method fails, for example when set *fin-zuinf* contains the verb *scheinen* because of a single occurrence of this verb in a misclassified context.

[12] These verbs are often called 'obligatorily coherent verbs', see [Haider 1993].

Therefore we decided to compare the frequency figures of the two sets rather than the verb sets themselves: a verb in *zuinf-fin* is a successful candidate, when its frequency in *freq-zuinf-fin* is high compared with its frequency in *freq-fin-zuinf*. This reflects the idea that a big difference between the two frequencies of a given verb indicates a tendency of this verb to prefer one context to the other. Provided that there is little noise in set *fin-zuinf*, the method of comparing the frequency distributions should work very well.

One way of implementing the comparison is to check for each verb in set *zuinf-fin* whether the quotient $\frac{\text{freq. in set } \textit{freq-fin-zuinf}+1}{\text{freq. in set } \textit{freq-zuinf-fin}+1}$ is suffciently small; as bias we experimentally chose 0,02. A lower value of the bias leads to more silence, whereas a higher value results in less silence, but possibly more noise: then we might also get verbs which actually do allow extraposition of the zu-infinitive, and which are simply not frequently used in the text corpus.

Linguistic queries. For the extraction of the two verb sets, two query templates have been designed, which are illustrated in Figures 3 and 4, respectively. The purpose of the query templates is to find verbs which subcategorize only for a subject and an infinitival complement with *zu*.

Fact	Encoding	Examples	
subord. conj.	[pos = "KOUS"]	als	weil
item sequence: no verb no punctuation no "es"; up to 12 items	 [POS ≠ "V.*" & POS ≠ "IP.*" & word ≠ "es"] {1,12}	 Maria gestern Birnen	 Hans heute Äpfel
"zu"	[word = "zu"]	zu	zu
infinitive	[pos = "V.INF"]	kaufen	verkaufen
finite verb	[pos = "VVFIN"]	**versuchte**	**scheint**
within a sentence	within s		

Figure 3: Query for set *zuinf-fin*: find verbs taking subject and infinitival complement with *zu* in contexts with the finite verb following the zu-infinitive.

The query template for set *zuinf-fin* is designed to collect as many verbs as possible. Therefore verb complements are modelled indirectly by excluding certain categories like verbs and punctuation marks. In this context (zu-infinitive precedes the finite verb), complements of the zu-infinitive can not be distinguished from complements of the finite verb by means of POS-shapes. Hence, the resulting verb set will possibly contain misclassified verbs, which take not only a subject and a zu-infinitive, but also a direct or indirect object.

The query template for set *fin-zuinf* on the other hand, is designed to avoid as much noise as possible. Here we try to identify only verbs taking subject and zu-infinitive but no other complements. We do this by explicitely modelling the subject noun phrase in the matrix clause with a complex NP-POS-shape[13]. Constraing the context in this way minimizes noise in *freq-fin-zuinf*, which is important for the subsequent comparison of the frequency distributions.

Fact	Encoding	Examples	
subord. conj.	[pos = "KOUS"]	als	weil
NP	complex NP-POS-shape	Maria	ein kleines Mädchen
adverbs	exclusion of non-adverb categories	gestern	heute
finite verb, but not: v. of ex.	[pos = "VVFIN" & !file(verbs-of-existence)]	**versuchte**	**versucht**
comma	","	,	,
no verb no punctuation no "es"	[POS ≠ "V.*" & POS ≠ "IP.*" & word ≠ "es"]	Birnen	Äpfel
"zu"	[word = "zu"]	zu	zu
infinitive	[pos = "V.INF"]	kaufen	verkaufen
within a sentence	within s		

Figure 4: Query for set *fin-zuinf*: find verbs taking subject and infinitival complement with *zu* in contexts with the zu-infinitive following the finite verb.

Results and Evaluation. We applied the query templates to a corpus of German newspaper text of about 200 million tokens. One important result of the extraction experiment is that there are only very few obligatorily coherent verbs, i.e. which do not allow extraposition of the zu-infinitive. After automatically comparing the frequency distributions of the two verb sets, a list of 11 verbs remained.

The list has been manually checked to identify and remove misclassified verbs and to test for the other verbs, whether extraposition of the zu-infinitive is possible[14]. From the remaining 8 verbs, only one, namely *verstehen*, actually does allow extraposition of the zu-infinitive. Figure 5 in annex A shows the tested verbs together with evidence phrases and (manually made-up) test phrases.

[13]Furthermore, as finite verbs we exclude verbs of existence like *bestehen, bleiben.* These verbs tend to occur with nouns taking zu-infinitives, such as *weil die* Möglichkeit *besteht, ein Buch zu lesen.*

[14]By looking at the automatically collected evidence phrases from set *zuinf-fin*, we found 3 misclassified verbs, which subcategorize not only for subject and zu-infinitive, but also for a direct or indirect an object.

4 Assessment

4.1 Fragment

The set of queries for subcategorization extraction is still under construction. As of early July 1996, the following types of constructions and their combinations can be extracted from German texts:[15]

- verbs with subject only (intransitive);
- verbs with subject and accusative and/or dative `NP` complement;
- verbs with subject (optional complement) and correlate construction (pointing to a prepositional object);
- verbs with subject (optional complement) and *zu*-infinitive, or complement clause with complementizer *daß* or *wh*-words.

Currently, some 3.000 verb readings have been extracted and validated. The noise rate is relatively low: on 1325 candidate verbs for the pattern "verb<[NP-NOM][NP-ACC]>", 57 items (= less than 5%) have been identified which do not qualify as dictionary-relevant.

4.2 Linguistic problems – possible solutions

The approach has a number of limitations, some of which are inherent to the use of a regular grammar. Moreover, the automatically tagged material contains the usual percentage of errors.

The procedures only allow to find the constructions we search for; the approach is dependent on the model of subcategorization classes used, and on the presence, for each class, of a discovery procedure.

A major limitation of the extraction devices is due to the use of a regular grammar: only sequences of phrase structural constructs can be identified, and no inference about grammatical functions is possible whenever the relationship between the pair of <phrasetype, case> and the grammatical function is not 1:1. Thus transitive (passivizable) verbs like *kaufen* and verbs taking a circumstantial complement (duration: *dauern* - *die Sitzung dauert eine Stunde*; weight: *wiegen* - *er wiegt 100 Kilogramm*; etc.) or an adverbial (*er arbeitet jeden Tag, er kauft eine Menge Waren*) are extracted by the same routine and need to be separated out manually[16].

[15] As stated above, we have to keep track of word order variation, active/passive, complex tense, and morphosyntactic properties of the verbs (reflexive, separable prefix, etc.), as well as of combinations of these.

[16] A subset of the passivizable verbs could be identified automatically: those actually occurring in the passive in the corpus. Verbs taking a "theme" and an "experiencer" (*Die Frage interessiert ihn*) are also in the noise set.

Similar problems, well known from any theoretical work on valency dictionaries, concern the distinction between indirect objects and free datives, and, between complement and adjunct prepositional phrases[17].

5 Future work

Current work is aimed at completing the fragment coverage. In addition, work on noun and adjective subcategorization has started. For example, material for prepositional attributes of nouns has been extracted (*Freude auf...*, *Interesse an...*, etc.).

In a parallel strand, the corpus exploration tools will be used to validate data from machine-readable dictionaries in text corpora: the subcategorization information contained in an electronic dictionary will be used to parameterize queries for individual verbs. For each subcategorization indication from the dictionary, corpus evidence will be sought. Dictionary indications not documentable with corpus data will be manually assessed.

Another important dimension to follow is some sort of semantic clustering of the results. The entry in figure 8 clearly shows the need for this, since it contains at least two readings of the verb, one as a speech act, and one as an abstract collocate (*das Feuer fordert ein Todesopfer.*) A combination of our approach with one that allows for statistical clustering of heads of verb subjects and complements seems most promising.

References

[Abney 1991] Steven Abney: "Parsing By Chunks", in: Robert Berwick, Steven Abney and Carol Tenny (Eds.): *Principle-Based Parsing*, (Dordrecht: Kluwer Academic Publishers), 1991

[Christ 1994a] Oliver Christ: "A Modular and Flexible Architecture for an Integrated Corpus Query System", in: Ferenc Kiefer, Gábor Kiss, Júlia Pajzs (eds.): *Papers in Computational Lexicography, COMPLEX '94*, Budapest, 1994, pp. 23-32.

[Christ 1994b] Oliver Christ: "The XKwic User Manual", internal report, Stuttgart: IMS, 1994.

[Grishman/MacLeod/Meyers 1994] Ralph Grishman, Catherine MacLeod, Adam Meyers: *Comlex Syntax: Building a Computational Lexicon*, (New York: New York University), 1994.

[Grishman/MacLeod 1994] Ralph Grishman, Catherine MacLeod: *COMLEX Syntax Reference Manual Version 1.1*, Draft prepared for the Linguistic Data Consortium, University of Pennsylvania, 1994.

[Haider 1993] Hubert Haider: *Deutsche Syntax - generativ: Vorstudien zur Theorie einer projektiven Grammatik*. Gunter Narr Verlag, Tübingen, 1993.

[17] See above, section 3.1. Frequency can help somewhat with PPs: not only the relative frequency of a single <[NP-NOM][PP]>-construction is relevant, but also the range of prepositions found along with a given verb, and the relative importance of the prepositions of each verb.

[Hindle 1991] Donald Hindle: "Structural Ambiguity and Lexical Relations", in: *Proceedings of the 29th Annual Meeting of the ACL*, 1991: 229-236

[Rooth/Carroll 1996] Mats Rooth, Glenn Carroll: "Valence Induction with a Head-Lexicalized CFG", (Stuttgart: IMS), ms. 1996

[Schiller/Teufel/Thielen 1995] Anne Schiller, Simone Teufel, Christine Stöckert, Christine Thielen: "Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS", Stuttgart/Tübingen, 1995.

[Schulze 1996] Bruno Maximilian Schulze: *MP user manual*, Stuttgart: IMS, 1996.

[Teufel 1995] Simone Teufel: *ELM-DE: A typed incarnation for German of the EAGLES Standard Proposal for Morphosyntactic Annotation – Lexical Specification and Classification Guidelines*, (Stuttgart/Pisa: IMS/EAGLES) 1995, ms. 172 pp. See also the electronic version, on the URL of the EAGLES project: `http://www.ilc.pi.cnr.it/EAGLES/home.html`

[Voutilainen et al. 1992] Atro Voutilainen, Juha Heikkilä, Atro Anttila: "Constraint Grammar of English: A Performance-Oriented Evaluation", Publication No. 21, University of Helsinki, Department of General Linguistics, 1992.

A Appendix: Data

Det/Pron	Case	Examples
Det: article	nom/acc	*das*, *die*. *ein*
Det: demonstr.	nom/acc	*diese*, *jene*, *derselbe*, *derjenige*, *dasjenige* ...
Det: indef.	nom/acc	*irgendein*, *irgendeine*, *alle*, *jede*, *kein*, *manche*, ...
Det: article	dat	*dem*, *einem*.
Det: demonstr.	dat	*diesem*, *jenem*, ...
Det: article	gen	*des*, *eines*.
Det: demonstr.	gen	*desselben*, *desjenigen*, ...
Pron: pers.	nom	*ich*, *du*, *wir*.
Pron: demonstr.	nom	*derselbe*, *derjenige*,
Pron: indef.	nom	*man*, *jemand*, *niemand*, *irgendwer*.
Pron: pers.	acc	*ihn*
Pron: demonstr.	acc	*den*,
Pron: indef.	acc	*irgendwen*, *jeden*, *niemanden*, ...

Table 1: Determiners and Pronouns with unambiguous morphosyntactic case forms

Verb	Evidence phrase	Test phrase
brauchen	`daß er nichts ernstzunehmen braucht`	⋆ daß er braucht, nichts ernstzunehmen
pflegen	`daß sie ihre Kritiker zu überleben pflegt`	? daß sie pflegt, ihre Kritiker zu überleben
scheinen	`obwohl er nach hinten zu kippen scheint`	⋆ obwohl er scheint, nach hinten zu kippen
suchen	`als ein Lkw eine Straßensperre zu durchbrechen suchte`	? als ein Lkw suchte, eine Straßensperre zu durchbrechen
trachten	`obwohl die Regierung dies zu verhindern trachtete`	? obwohl die Regierung trachtete, dies zu verhindern
vermögen	`daß man etwas zu leisten vermag`	? daß man vermag, etwas zu leisten
verstehen	`der mit Sprache umzugehen versteht`	der versteht, mit Sprache umzugehen
wissen	`daß sie Risiken abzuschätzen weiß`	? daß sie weiß, Risiken abzuschätzen

Figure 5: The tested verb candidates from the resulting verb list.

B Examples of results

sehen	463
kennen	380
wissen	344
machen	269
brauchen	244
tun	187
finden	147
verstehen	145

Figure 6: Part of the frequency distribution of verb lemmas in 200 million words, with the subcategorization pattern *verb*<[NP-NOM][NP-ACC]>

Die überschätzen ihre eigene Kraft sehr .
die Bahn übersehe hier die topographische Lage .
Die Sensoren übersetzen die Bewegungen .
Ein Dolmetscher übersetzte die Vernehmung .
Das Frisierstübchen übersieht man fast .
Die Bilanzsumme übersprang die Acht-Milliarden-Mark-Grenze .
Er überstand die Vertrauensabstimmung unbeschadet .
Die Inszenierung übersteht gerade mal die Premiere .
Die gemessene Radioaktivität übersteige nicht die zulässige Norm .

Figure 7: Sentences illustrating verbs prefixed with "*über*" which subcategorize for a nominative and an accusative NP

< record > < verb > fordern < /verb >
< subcat > subj(NP_nom) obj(NP_akk) < /subcat >
< typical >
Aber niemand fordert ihre Legalisierung .
Auch die Seeleutegewerkschaft fordert ein umfassendes Waffentransportverbot .
Auch die afghanische Nachbarregierung fordert ihre Freilassung .
Das Büro fordert nun die Rekonstruktion des Kunstwerks .
Das Feuer fordert ein Todesopfer :
Das fordere auch nicht das Sozialstaatsprinzip .
Das fordern die niedersächsischen Christdemokraten .
< /typical >
< /record >

Figure 8: A sample proto-entry for *fordern*<[NP-NOM][NP-ACC]>