

# Lexicon Acquisition with and for Symbolic NLP-Systems – a Bootstrapping Approach

Jonas Kuhn, Judith Eckle-Kohler & Christian Rohrer

Institut für maschinelle Sprachverarbeitung  
Universität Stuttgart  
Azenbergstraße 12  
70174 Stuttgart  
{kuhn,eckle,rohrer}@ims.uni-stuttgart.de

## Abstract

We present a method of applying a broad-coverage LFG grammar of German in the process of semi-automatic lexicon acquisition from corpora. The identification of corpus instances that illustrate a certain subcategorization frame uniquely is done by a comparison of the numbers of analyses the grammar assigns to the corpus instances, under the assumption of different hypothetical lexicon entries for the candidate verb. Filtering conditions expressed on the feature representation output by the grammar further restrict the sentences that the automatic extraction step is based on. Experiments show that the grammar-based method produces better results than a method based on patterns in a corpus query language.

## 1. Background

This paper reports ongoing research activities in methods for semi-automatic lexicon acquisition from corpora (cf. also (Eckle and Heid, 1996; Eckle-Kohler, 1998)). As a test application, the lexical resources constructed with various methods are being used in a broad-coverage LFG<sup>1</sup> grammar of German under development at the IMS Stuttgart. With the method reported in this paper, a bootstrapping cycle is closed: the lexical resources are no longer just *applied* in the LFG grammar, but application of the grammar also feeds back into the construction of further resources.

The grammar development activities are part of a research project on grammar engineering and the Parallel Grammar Project<sup>2</sup> (in collaboration with Xerox PARC and the Xerox Research Centre Europe).

Acquisition of subcategorization lexicons is an important factor in the construction of broad-coverage symbolic grammars—the interaction of lexical resources in the German LFG grammar is organized according to the following considerations: Using the Xerox Linguistic Environment (XLE)<sup>3</sup> as the grammar development platform and parsing system, the grammar interfaces with a two-level morphological analyzer for German (Schiller, 1994), including a lexicon with more than 40.000 stems (across all categories). For word classes with a trivial subcategorization behaviour (like common nouns and many adjectives), the grammar can exploit the categorization already carried out in the morphological analyzer; i.e., all the syntactic lexicon has to provide for these classes are generic entries to be assigned to any lemma the analyzer recognizes as a member of this class. However, for verbs (and for relational adjectives and nouns), an idiosyncratic assignment of subcategorization frames for each lemma is indispensable if the grammar is

supposed to produce accurate and deep analyses, as they are required, e.g., in high-quality machine translation. Now, unless the domain to be covered by the system is restricted radically, it is impractical to hand-encode the required subcategorization lexicon; thus the success of broad-coverage symbolic parsing depends to a considerable degree on ways of automating the construction of high-quality lexical resources.

## 2. Problem and Methodology

Currently for German (like for most if not all other languages), no sufficient amount of syntactically analyzed corpora (tree banks) is available as an input for lexicon acquisition. Thus, part-of-speech tagged corpora have to suffice as the basis for the acquisition process. The problem is that in the general case, a corpus instance for a candidate lemma will allow no unique conclusion about the subcategorization frame of the lemma; e.g., morphological case marking is often ambiguous, or a *that*-clause may be an argument either of the verb or of some noun.

### 2.1. Semi-automatic lexicon acquisition

While in stochastic NLP systems, inaccuracies originating from misclassifications in fully automatic lexicon acquisition can be expected to be minimized by low probabilities, such inaccuracies can have drastic negative effects in a symbolic system. Thus, resources to be used in such a system should at least undergo a manual check. The ideal acquisition method is a semi-automatic one with very high precision in the automatic phases combined with an efficient scheme of post-editing by a human lexicographer (Eckle-Kohler, 1998). This can be achieved by relying as much as possible on corpus instances that indicate a certain syntactic property unambiguously. Provided high precision is reached and an efficient post-editing is in fact possible, the relatively low recall—due to the non-ambiguity requirement—can be compensated by the use of large corpora.

<sup>1</sup>Lexical-Functional Grammar (Kaplan and Bresnan, 1982; Dalrymple *et al.*, 1995).

<sup>2</sup>See <http://www.parc.xerox.com/istl/groups/nltt/pargram/>

<sup>3</sup>On lexical resource reconciliation in XLE, cf. (Kaplan and Newman, 1997).

## 2.2. Alternatives in Implementation

Eckle-Kohler (1998) discusses a successful implementation of the mentioned semi-automatic acquisition strategy, applying patterns in the corpus query language CQP (Schulze, 1996), combined with a number of postfiltering steps, to identify unique corpus instances of a particular linguistic property (e.g., membership in a particular subcategorization class). This method is in effect a simulation of a shallow parser with a special-purpose grammar for a (nearly) unambiguous pattern that instantiates the property under consideration. All that the manual post-editing amounts to is a verification that the candidate lemmata extracted for a certain frame can be generally used with this frame, in order to detect “false positives” (Briscoe and Carroll, 1997). In the present paper, we investigate an alternative approach to arrive at a selection of corpus instances of a candidate lemma that are unique with respect to the range of potential linguistic properties under consideration (or, more concretely for the subcategorization frame extraction task: a selection of corpus instances that uniquely illustrate a particular frame). The basic idea is to exploit the independently developed LFG grammar—including all previously acquired lexical knowledge—to identify unique instances. A set of contexts containing the candidate lemma is extracted from a corpus and a first prefiltering step is applied that ensures that some basic structural requirements are met (e.g., only subordinate clauses with the finite verb in final position are allowed). These contexts are then parsed a number of times, each time stipulating a particular subcategorization frame for the unknown verb candidate. A comparison of the successful parses reveals which instances are unambiguous and thus a suitable base for conclusions about the actual properties of the candidate.

Some contexts will receive an analysis for more than one hypothetical entry. Sentence (1) is an example of this type of contexts.

- (1) weil unser Präsident Maier halluziniert, daß es  
*because our president M. hallucinates that it*  
schneit.  
*snows*

Assume the task is to identify verbs subcategorizing for a subject and a finite object clause as opposed to verbs taking an additional object NP (the task discussed in detail below). Although the usage of the verb *halluzinieren* ‘hallucinate’ in (1) is actually an instance of the former frame, the sentence could just as well be analyzed if the second frame was correct: in this case the string *unser Präsident Maier* is considered to consist of two NPs, where the latter—*Maier*—can be accusative or dative. Had the verb been *vorwerfen* ‘reproach’ rather than *halluzinieren*, this would have been the correct analysis.

Thus, examples of type (1) are no suitable basis for automatic conclusions about the actual subcategorization frame. They are detected since the grammar produces an analysis for two hypothetical frames, and will be thrown away. Conforming to the general strategy of semi-automatic lexicon acquisition (cf. sec. 2.1), the idea is as follows: if there are sufficiently many occurrences of a lemma in a corpus,

there will be enough unambiguous ones, such that the unclear ones can be simply thrown away.

The main difference between (i) the original query-based method of (Eckle-Kohler, 1998) and (ii) the new method just sketched is thus a conceptual one: at which point of the process is uniqueness of the instances enforced, and where is the knowledge about the morphosyntactic properties of unique instances encoded? In (i), uniqueness is enforced by an appropriate formulation of the query patterns and the postfiltering step; i.e., the knowledge which surface properties guarantee uniqueness is explicitly encoded. For the alternative method (ii), the same is certainly true for the prefiltering step (and an additional grammar-based postfiltering discussed in sec. 3.2 below); however, the key step, based on the hypothetical lexicon entries, exploits the fact that knowledge about surface properties of unique instances is already implicitly encoded in the grammar. This knowledge can be brought to the surface by disjunctively presenting the full range of possibilities (i.e., all potential frames), applying them to a particular instance and filtering out non-exclusive cases (i.e., instances that receive a solution for more than one hypothetical frame). The main design task thus lies in determining which options (frames) should go into the range of possibilities, and in singling out (by explicit pre- and postfiltering) contexts in which a choice on the basis of this range can be performed with minimal distortion.

The present paper establishes the basic methodology of the grammar-based acquisition method (ii) and attempts to assess its quality by comparing it with the original method (i) within a domain where the latter has been applied very successfully—the choice between a particular range of subcategorization frames for a candidate verb.

## 3. Task

We illustrate and assess the proposed acquisition technique for a particular task: the extraction of German verbs subcategorizing for a finite argument clause (the equivalent of a *that*-clause) from a 200 million token newspaper corpus. The experiments aim at the extraction of three different subcategorization frames involving finite object clauses; the following examples illustrate the frames (abbreviations are given in the obvious way, listing the categories of subcategorized elements with relevant additional information—case and form of complementizer):

- (2) NPnom-Sdass  
Otto vermutet, daß es schneit.  
*O. suspects that it snows*
- (3) NPnom-NPdat-Sdass  
Anna verrät dem Kind, daß sie der Osterhase  
*A. reveals (to) the child that she the Easter bunny*  
war.  
*was*
- (4) NPnom-NPacc-Sdass  
Maria überzeugt den Studenten, daß er kommen muß.  
*M. convinced the student that he come must*

The NPnom-Sdass and NPnom-NPdat-Sdass classes are very large; many verbs can appear with either of the two frames, either because the dative NP is an optional argument (e.g., in *glauben* ‘believe’, *versprechen* ‘promise’), or because there are two word senses with different numbers of argument (e.g., in *anmerken*—without NPdat: ‘note, say’; with NPdat: ‘notice s.th. at s.o.’). Verbs with the NPnom-NPacc-Sdass frame form a fairly small class. Finite argument clauses also appear in other subcategorization frames: with inherent reflexives (5), together with a pronominal adverb like *daran* in the place of a PP (6), or as subject clauses (7). In the experiment under discussion, it is not intended to extract instances for these frames; rather, it has to be avoided that such contexts come in as noise.

- (5) Peter bemüht sich, daß er rechtzeitig ankommt.  
*P. endeavours REFL that he in time arrives*
- (6) Anna denkt daran, daß sie versprochen hat zu kommen.  
*A. thinks of it that she promised has to come*
- (7) Seit gestern steht fest, daß er kommt.  
*Since yesterday is certain that he comes*

Further occurrences of *daß*-clauses that may lead to misclassifications are argument clauses of nouns in extraposed position (8), and adverbial *daß*-clauses (mostly with a correlative *so* in the matrix clause (9)).

- (8) Maria hat die Absprache durchbrochen, daß niemand das Haus erwähnt.  
*M. has the arrangement broken that nobody the house mentions*
- (9) Otto hat so laut gelacht, daß die Nachbarn aufgewacht sind.  
*O. has so loudly laughed that the neighbours woken up have*

## 4. Experiments

We tackled the extraction task in three different experiments. As a reference for evaluating the new grammar-based method, an experiment with the “query-based” method (as discussed extensively in Eckle-Kohler (1998) for a very similar extraction task) was performed; for the grammar-based method, we report two experiments that differ in the amount of pre- and postfiltering that was performed.

A certain amount of prefiltering was invariantly applied in all three experiments: the sentence structure to be analyzed is restricted to the “verb-last” structure of German subordinate clauses (which is linguistically considered the underlying structure), with the *daß*-clause following the verbal complex. This was achieved by extracting from the corpus contexts that start with an arbitrary subordinating conjunction excluding *daß* and that contain no punctuation indicating clause boundaries, apart from the commas pre-

ceding the *daß*.<sup>4</sup> Since the internal structure of the argument clause is irrelevant for the subcategorization properties of the matrix verb under consideration, the *daß*-clauses were uniformly replaced by a “dummy clause” (*daß er kommt*).

Further prefiltering removed contexts containing (a) nouns known (from other extraction experiments) to subcategorize for a *daß*-clause themselves (cf. (8)), (b) adverbs and adjectives which can function as a correlative of an adverbial *daß*-clause, like for example *so* (cf. (9)), and (c) pronominal adverbs which can also function as a correlative of a *daß*-clause, indicating its function as a prepositional object (cf. (6)).<sup>5</sup> Contexts containing the third person reflexive pronoun (*sich*) were also filtered out, since its usage with a verb that is not inherently reflexive will be generally indistinguishable from the inherently reflexive case (cf. (5)).<sup>6</sup> Finally contexts containing the pronoun *es* were filtered out, because in German this pronoun frequently indicates subject clauses.

### 4.1. Query-based reference experiment

In the query-based extraction method, CQP expressions are formulated to identify unique contexts for each of the frames under consideration.

The queries for the three frames NPnom-Sdass, NPnom-NPdat-Sdass and NPnom-NPacc-Sdass cover verb-last clauses in the active voice, including all possible constructions with auxiliaries and modals, which contain one lexical verb being considered as the candidate verb. Within a verb-last clause, each query pattern identifies the noun chunks, the verb cluster, as well as several optional adverbial and prepositional chunks.

The following constraints are shared by all three queries (they all exclude ambiguous structures that may lead to misclassifications): verb forms where the present perfect tense is formed with the auxiliary *sein* must not occur, any post-nominal genitive attribute of a noun has to be unambiguously genitive, and prepositional chunks where the preposition is subcategorized for by verbs are excluded.

The queries for the three frames are different with respect to the noun chunks they match in a verb-last clause: the query for the NPnom-Sdass frame matches a noun phrase containing either a common noun as its head or a pronoun; in the former case, the noun phrase consists either of an obligatory determiner which is unambiguously either nominative or accusative<sup>7</sup>, and which is followed by optional adjectives or adverbs, a common noun and an optional genitive attribute; in the latter case, it consists of a pronoun which is unambiguously either nominative or accusative.

<sup>4</sup>The upper limit of tokens between the conjunction and the comma in front of the *daß* was (arbitrarily) specified to be 26.

<sup>5</sup>In principle, all this filtering could be done with the grammar-based technique as well; but for the experimental purposes of this paper, it seemed appropriate to vary the technique for just a limited number of filtering steps.

<sup>6</sup>See also the discussion of first and second person reflexives in sec. 4.3.2.

<sup>7</sup>Unambiguous nominative determiners are rare, and often there is a nominative/accusative ambiguity which causes no major problem for the identification of the NPnom-Sdass frame, but excludes the more frequent subject clauses with dative NP.

The query for the NPnom-NPacc-Sdass frame matches clauses containing two such nominative/accusative NPs. The query for the NPnom-NPdat-Sdass frame matches clauses where the dative NP follows the subject NP. The subject NP differs from the nominative/accusative NP described above in two aspects: first the determiner is optional and secondly the determiner or pronoun is not constrained with respect to case. The dative NP is completely analogous to the nominative/accusative NP; the determiner or pronoun has to be unambiguously dative.

## 4.2. Initial grammar-based experiment

As mentioned in sec. 2.2, the basic idea of the grammar-based method is to compare how different hypothetical subcategorization frames for the candidate verb influence the parsability of a certain context from the corpus. In order to be able to apply this strategy, the candidate verbs have to be known before the parsing task is started, such that the hypothetical lexicon entries can be constructed appropriately. In the first grammar-based experiment, a relatively simple method is applied to construct the various hypothetical lexicon entries: like in the previous section, for each prefiltered sentence from the corpus, a single full verb in the verb cluster is identified as the candidate, using a query language expression. For the list of all candidates' stem forms, hypothetical lexicons are constructed for each of the possible frames, i.e., in our case the three frames NPnom-Sdass, NPnom-NPdat-Sdass and NPnom-NPacc-Sdass.

Other than for the mentioned general prefiltering and a restriction to active verb forms, no more query language expressions are used to narrow down the contexts to unique ones containing just the arguments of the subcategorization frame plus some optional modifiers. Instead, the independently developed LFG grammar is used for this purpose.

For the actual grammar-based part of the experiment, the grammar's ordinary lexicon for verb stems (interacting with the morphological analyzer) is replaced by the hypothetical stem lexicons.

The prefiltered contexts are presented in the ordinary XLE test suite format, such that the parser can process them in batch mode. Since all sentences have the structure of subordinate clauses, the root symbol is modified appropriately, reducing the search space. The contexts are parsed three times, once with each of the different hypothetical lexicons, recording the respective number of solutions produced for each sentence. The results from the three parsing cycles are thus filtered (by scripts) in such a way that only on the basis of unique contexts (as discussed in sec. 2.2), verb candidates can get into the list passed on to the lexicographer for manual checking.

The simple set-up of the initial grammar-based experiment is problematic: the interpretation of the parsing results, purely based on the numbers of solutions, presupposes that all the sentences getting a parse have a structure in which it is actually the candidate verb and only the candidate verb that is the head of the arguments under consideration. However, on the basis of the little prefiltering that is performed, this is not guaranteed. While the query patterns of sec. 4.1 will filter out all contexts that contain more than one full verb (due to coordination or raising and control verbs), they

are not generally excluded by the grammar, and caused distortion in the result.

A possible reaction would have been to add in additional prefiltering to enforce a more restricted structure for the contexts that go into the grammar experiment; however, this would have been against the spirit of the grammar-based method. We chose instead to make more use of the syntactic representation that the grammar assigns to the sentences analyzed: certain conditions on the representation form a postfilter restricting the set of parsed sentences that go into the comparison.

## 4.3. Final grammar-based experiment

The final grammar-based extraction experiment differs in three major ways from the initial one: Firstly, interpretation of the results no longer relies on the mere number of solutions obtained for a sentence under a particular hypothetical lexicon entry—rather, additional constraints on adequate output representations are expressed in order to filter the contexts that go into lexicon acquisition (cf. sec. 4.3.2). Secondly, the hypothetical lexicon entries are no longer stored in a single list for the whole of candidates, but are dynamically created for the individual candidate under current consideration; this conforms much more to the bootstrapping strategy, since other (non-candidate) verbs contained in a certain context are unaffected by the experiment-specific lexicon set-up, i.e., if knowledge about them was already acquired in earlier experiments, this knowledge will have positive influence on the current experiment (cf. sec. 4.3.1).

Furthermore, other than in the initial experiment, this time all full verbs occurring in the sentences extracted from the corpus were treated as a candidate (duplicating sentences with more than one candidate).

### 4.3.1. Dynamic lexicon adaption

Creating a special lexicon set-up for each individual candidate verb (in the experiment, there are 650 different candidates) under every hypothetical frame seems a very time-consuming task. However, the XLE system allows a runtime modification of the lexicon specification and provides a notation to overwrite individual entries with other information (the lexicon edit operators described by Kaplan and Newman (1997)). Thus, in principle the complete experiment could be run loading the XLE system and the LFG grammar just a single time. A script forms groups of corpus instances with the same candidate, and creates a new lexicon entry for the candidate verb stem, assigning the first hypothetical subcategorization frame and overwriting other potential entries for the same stem. With this lexicon entry active, the group of corpus instances is parsed, recording the number of solutions assigned. Next, a lexicon entry assigning the second hypothetical frame is created, overwriting the first one, etc.

### 4.3.2. Filter on output representations

Since in the grammar-based extraction method, deep linguistic analyses (in our case, lexical-functional syntactic analyses) are assigned to the corpus instances parsed, detailed structural information can be exploited to formulate

filters on what counts as an instantiation of the frame in question. Technically, this is done by the “TASTE” method, developed for **T**arget **S**pecification and regression **T**esting in grammar development, and described in (Kuhn, 1998). The idea is to augment the actual grammar with technically motivated rules, using a new root category. The productions for the new category map to the original category (plus, in the regression testing scenario, to additional test expressions). The purpose of the additional rules is to introduce conditions on the feature structure that the grammar produces for the root symbol. Here is an extremely simple example:

$$(10) \text{ ROOT}' \rightarrow \text{ ROOT} \\ (\downarrow \text{PASSIVE}) \neq +$$

Assuming that ROOT is the root symbol for the ordinary grammar, this rule filters out all analyses that have a + as the value of the feature PASSIVE in the feature structure corresponding to the ROOT node (thus the metavariable  $\downarrow$ ). Furthermore, application of the XLE constraint ranking mechanism of (Frank *et al.*, to appear) makes it possible to introduce a special *ungrammaticality* mark for certain disjuncts in constraints. The effect of this mark is if for a sentence the only solutions found involve such marked disjuncts, the parser puts an asterisk in front of the number it returns.

For a grammar including such marked disjuncts, there are three classes of sentences: those that cannot be analyzed at all by the rules plus feature annotations (receiving 0 solutions), those that have at least one unmarked analysis (receiving  $n$  solutions,  $n \geq 1$ ), and those with just marked analyses (receiving  $*n$  solutions,  $n \geq 1$ ).

The marks can now be applied to mark unwanted configurations in the filter of the extraction experiment. So, if we want to rule out passives, we can specify the following disjunctive condition:<sup>8</sup>

$$(11) \text{ ROOT}' \rightarrow \text{ ROOT} \\ \left\{ \begin{array}{l} (\downarrow \text{PASSIVE}) \neq + \\ (\downarrow \text{PASSIVE}) = + \\ \text{Ungrammatical} \in o* \end{array} \right\}$$

Now, sentences for which the grammar under ROOT produces an analysis with the feature value PASSIVE + will be distinguishable from sentences that are not covered by the ROOT-grammar at all. In the automatic extraction reported here, no direct use of this option was made; but it has been very valuable in fine-tuning the filter expressions, and may be used if for low-frequency words, filtering conditions should be relaxed to increase recall.

With the method just introduced, the following three filter conditions were expressed in the actual experiment:

1. Personal pronouns of first and second person were disallowed as objects, if the subject was in the same person. The reason is that in such a context, there is no way of telling whether the verb is inherently reflexive (as in (12)) or not.

<sup>8</sup>Technically, marks like ‘Ungrammatical’ are introduced in a special *o*-projection, which is set-valued, thus the membership statement.

(12) weil ich mich gefreut habe, daß ich Arbeit  
because I REFL please have that I work  
gekriegt habe  
got have  
‘because I was pleased, that I got a job’

With a subject in a different person, like in (13), it is however clear that the object has to be an ordinary thematic argument, so they can be used for lexicon acquisition.

(13) wenn mich jemand überzeugen kann, daß diese  
if me anybody convince can that these  
Sanktionen nur das Regime treffen  
sanctions only the regime hurt  
‘if anybody can convince me that ...’

It is one of the merits of the grammar-based method that such a condition can be expressed. Here is the actual condition in XLE notation:<sup>9</sup>

```
{ { (!SUBJ PERS) = 1
  | (!SUBJ PERS) = 2 }

{ ~(!OBJ)
  | (!OBJ PERS) ~= (!SUBJ PERS)
  | (!OBJ PERS) = (!SUBJ PERS)
  Ungrammatical $ o::* }

{ ~(!OBJ2)
  | (!OBJ2 PERS) ~= (!SUBJ PERS)
  | (!OBJ2 PERS) = (!SUBJ PERS)
  Ungrammatical $ o::* }

| (!SUBJ PERS) ~= 1
| (!SUBJ PERS) ~= 2
}
```

2. The *Zustandspassiv* (stative passive) formed with the auxiliary *be* and the past participle is excluded, since the combination may as well be a present perfect form (in active voice, thus suggesting a different subcategorization frame):

(14) weil die Kontrahenten übereingekommen/  
because the opponents agreed/  
überrascht sind, daß es eine Lösung geben muß  
surprised are that there a solution be must  
‘because the opponents have agreed/are surprised that there must be a solution’

Other passives are allowed, including the ‘modal infinitive’ as in the following example, which is a perfectly unique instantiation for the (passivized) NPnom-NPdat-Sdass frame:<sup>10</sup>

<sup>9</sup>In this notation, ‘!’ replaces the ‘ $\downarrow$ ’, ‘~’ expresses negation, ‘{ X | Y }’ the disjunction of X and Y; and membership in the *o*-projection is expressed as ‘\$ o::\*’. The condition can be paraphrased as follows: if the SUBJ is first or second person, then for both the two functions OBJ and OBJ2, the following is true: either (i) this function doesn’t exist, or (ii), it’s of a different person than the SUBJ, or (iii), it’s of the same person, in which case an ‘Ungrammatical’ mark gets introduced. Finally, if the SUBJ is neither first nor second person, nothing needs to be said (recall that the third person reflexive has already been filtered out in the preprocessing step).

<sup>10</sup>The active voice paraphrase in German would be as follows:

(15) weil der älteren anzusehen ist, daß sie gelebt  
*because the-DAT older to-look-at is that she lived*  
 hat  
*has*

‘because one can tell by looking at the older one that she’s lived her life’

3. Coordination of two verbs was disallowed, unless both conjuncts contained their own subject and the candidate forms the rightmost conjunct. This condition excludes contexts like the following:

(16) ... damit sie sie anrufe und ausrichte, daß er noch  
*so that she her phone and tell that he still*  
 lebt.  
*lives*

‘so that she’d phone her up and tell her, that he’s still alive’

Since in such a coordination, it is unclear which arguments are distributed over both conjuncts it may lead to misclassifications. Other coordinations, e.g., within NPs, are allowed.

## 5. Results

Evaluation focuses mainly on the final grammar-based experiment, and how its results compare with the query-based reference experiment.

The grammar-based method would have been problematic if the coverage of the symbolic grammar had not been large enough to be able to parse a significant part of the candidate sentences. The grammar was not particularly adapted to the newspaper corpus, since the main line of grammar development is focused on technical documentation texts. Given this background, the parsing rate was surprisingly high: in the 5139 contexts, there were just 1395 (27.1%) that the grammar was not able to analyze with any of the frames.<sup>11</sup> These include ill-formed strings from the corpus (e.g., mistaken to be verb-last clauses due to tagging errors) and sentences, in which the candidate verb is used with an infinitival complement—a frame that was not presented among the hypotheses.

The rate shows that in a clear-cut task (e.g., with the restriction to verb-last clauses), even a symbolic grammar without special robustness devices can be effectively applied.

### 5.1. Evaluation

Each experiment produces three lists of verbs: the candidates it suggests for each of the three subcategorization frames (that will go into the step of manual assessment by the lexicographer).

(i) weil man der älteren ansehen konnte, daß ...  
*because one the-DAT older look-at could that*

<sup>11</sup>The parsing of a sentence was aborted after a time limit of 60 seconds; most sentences were processed in less than a second each.

A precision measure was computed for all three experiments. It reflects the percentage of true positives (i.e., correct hypotheses) among all the verb candidates. The tables in fig. 1 give these measures, plus the absolute number of true positives extracted (#TP).

Query-based experiment		
frame	precision	#TP
NPnom-Sdass	92.0%	184
NPnom-NPdat-Sdass	87.8%	36
NPnom-NPacc-Sdass	9.5%	2

  

Initial grammar-based experiment		
frame	precision	#TP
NPnom-Sdass	82.4%	239
NPnom-NPdat-Sdass	74.1%	60
NPnom-NPacc-Sdass	7.3%	6

  

Final grammar-based experiment		
frame	precision	#TP
NPnom-Sdass	90.2%	193
NPnom-NPdat-Sdass	87.8%	58
NPnom-NPacc-Sdass	10.9%	7

Figure 1: Precision measures

The numbers for the initial grammar-based experiments reflect the underconstrainedness already discussed in sec. 4.2. The absolute number of true positives could be increased towards the query-based reference experiment, but only at the cost of dropped precision. The additional grammar-based postfiltering of the final experiment brought back the precision of the reference experiment, still with a greater number of successfully extracted true positives.

This gain in recall is more clearly reflected by the following measure taking into account the individual instances of a number of candidate verbs in the corpus: six verbs with a medium to low frequency in the extracted verb-last clauses were chosen, and for each occurrence in the verb-last clauses the actual subcategorization frame was manually assessed and annotated.<sup>12</sup> Verbs that can be used both with the NPnom-Sdass frame and the NPnom-NPdat-Sdass frame were chosen; each of the frames was evaluated individually.

The recall value for a verb used with a particular frame is the percentage of contexts exploited by the respective automatic method among all the manually identified contexts. The table in fig. 2 lists the measures for the query-based reference experiment (cf. sec. 4.1) and the final grammar-based experiment (cf. sec. 4.3). In the second column the number of corpus instances with that frame (as manually assessed) is given.

<sup>12</sup>The motivation for choosing medium to low frequency verbs is on the one hand, that this reduces the number of sentences which have to be manually annotated; but also on the other hand, that the recall values for medium to low frequent verbs indicate the capability of the acquisition system to be successfully applied to (usually small) sublanguage corpora.

<i>verb candidate</i> used with frame	actual occurr.	recall in exper.:	
		qu.-b.	gr.-b.
<i>versichern</i>			
NPnom-Sdass	23	13.0%	47.8%
NPnom-NPdat-Sdass	16	25.0%	75.0%
<i>ankündigen</i>			
NPnom-Sdass	27	22.2%	51.8%
NPnom-NPdat-Sdass	1	100%	100%
<i>bekanntgeben</i>			
NPnom-Sdass	19	31.6%	57.9%
NPnom-NPdat-Sdass	0	–	–
<i>versprechen</i>			
NPnom-Sdass	7	0%	57.1%
NPnom-NPdat-Sdass	11	27.3%	54.5%
<i>anmerken</i>			
NPnom-Sdass	8	25.0%	62.5%
NPnom-NPdat-Sdass	2	50.0%	100%
<i>zugestehen</i>			
NPnom-Sdass	5	60.0%	60.0%
NPnom-NPdat-Sdass	3	66.7%	66.7%

Figure 2: Recall measures

Throughout the examples, the recall of the reference experiment could be kept the same or was considerably improved in the grammar-based experiment. This means that in particular for low-frequency verbs or for small corpora (corpora for sublanguages), the grammar-based method can be expected to return better results than the purely query-based.

## 6. Conclusion

We reported a bootstrapping approach of lexicon acquisition with and for a symbolic grammar, making essential use of the depth of analysis of the grammar in automatic extraction of knowledge from corpora: the identification of corpus instances that illustrate a certain subcategorization frame uniquely is done by a comparison of the numbers of analyses the grammar assigns to the corpus instances, under the assumption of different hypothetical lexicon entries for the candidate verb. Filtering conditions expressed on the feature representation output by the grammar further restrict the sentences that the automatic extraction step is based on. Experiments show that the grammar-based method produces better results than a method based on patterns in a corpus query language.

## 7. Acknowledgements

The research reported in this project was in part funded by the *Deutsche Forschungsgemeinschaft* within the *Sonderforschungsbereich 340*, project B12 (Methods for extending, maintaining and optimizing a comprehensive grammar of German).

- Briscoe, T. and Carroll, J. (1997). Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, Washington, DC, USA.
- Dalrymple, M., Kaplan, R. M., Maxwell, J. T., and Zaenen, A., eds. (1995). *Formal Issues in Lexical-Functional Grammar*. CSLI Publications, Stanford, CA.
- Eckle, J. and Heid, U. (1996). Extracting raw material for a German subcategorization lexicon from newspaper text. In *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX '96*, Budapest.
- Eckle-Kohler, J. (1998). Methods for quality assurance in semi-automatic lexicon acquisition from corpora. In *Proceedings of EURALEX*.
- Frank, A., King, T., Kuhn, J., and Maxwell, J. (to appear). Optimality Theory style constraint ranking in large-scale LFG grammars. In *Proceedings of LFG98*, CSLI Online Publications.
- Kaplan, R. M. and Bresnan, J. W. (1982). Lexical-Functional Grammar: a formal system for grammatical representation. In J. W. Bresnan, ed., *The Mental Representation of Grammatical Relations*, ch. 4, 173–281. MIT Press, Cambridge, MA.
- Kaplan, R. M. and Newman, P. S. (1997). Lexical Resource Reconciliation in the Xerox Linguistic Environment. In D. Estival, A. Lavelli, K. Netter, and F. Pianesi, eds., *Computational Environments for Grammar Development and Linguistic Engineering, ACL Workshop*, 54–61, Madrid.
- Kuhn, J. (1998). Towards data-intensive testing and applications of a broad-coverage LFG grammar—partial target specifications as a filter on parser output. Ms., Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.
- Schiller, A. (1994). Deutsche Flexions- und Kompositions-morphologie mit PC-KIMMO. In *LDV-Forum – Forum der Gesellschaft für Linguistische Datenverarbeitung GLDV*, vol. 11 (1).
- Schulze, B. M. (1996). MP user manual. Ms., Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.