IMSLex – Representing Morphological and Syntactic Information in a Relational Database

Wolfgang Lezius, Stefanie Dipper, Arne Fitschen
IMS, University of Stuttgart

Abstract

We present a lexical resource comprising morphological and syntactic information. The resource is realised as a relational database, which facilitates the access and administration of the data. Sophisticated tools have been developed to allow a user-friendly usage of the resource. One application, a broad coverage parser, which makes use of both the morphological and syntactic part of the database, is presented in detail.

1 Introduction

In the last years, several lexical resources have been developed at the Institute for Natural Language Processing (IMS) of the University of Stuttgart. These include the lexicon of a morphology tool ([Schiller 1996]), a tagger lexicon ([Schmid 1994]), a lexicon for subcategorisation information ([Eckle 1999]), and special lexicons used for syntactic parsing ([Bröker/Dipper 1999]). The purpose of IMSLex is to link together these resources to build a common one. Thus, different levels of linguistic description are accessible in a unique source. This resource is realised as a relational database comprising tables corresponding to the individual (level-specific) sublexicons. Information from the different sublexicons is linked through a common table containing lemma information and pointers to the readings relevant for each level.
The main advantage of a database approach is the facility to access centrally stored data which is always up to date. Besides, the access and manipulation of the resource is realised via SQL, a powerful standardised query language. SQL programming interfaces are available for all commonly used programming languages, which enables easy integration of the lexicons into applications.

The architecture of IMSLex is illustrated in figure 1: the resource will be extended continuously by using (semi-)automatic acquisition from large text corpora and, as a second strand, an interactive maintenance tool (LexiTool). The application lexicons can be derived from the database. Currently, the resource is used by NLP systems for tagging, morphological analysis and broad coverage parsing in the framework of Lexical Functional Grammar (LFG). To query the database we have developed a sophisticated graphical user interface (LexiExplorer).
In the following two sections, the morphological and syntactic parts of the architecture are described in detail (cf. section 2 and 3, respectively). Finally, we illustrate how the lexicons together feed the parsing application (cf. section 4).

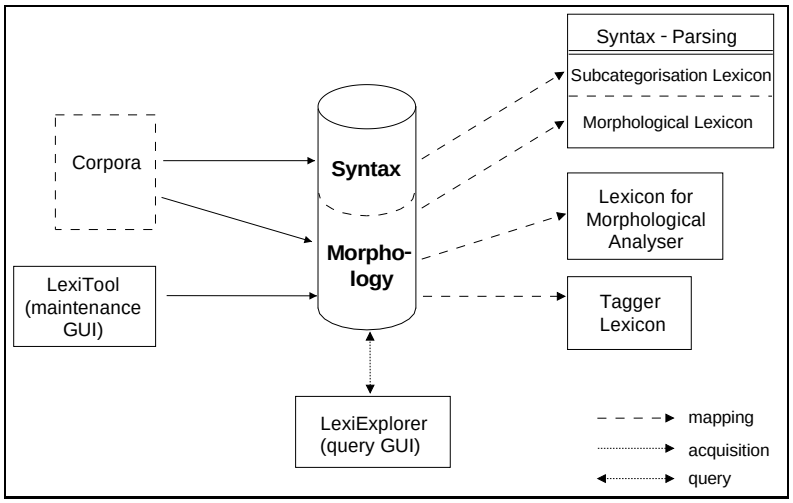

Figure 1: Architecture of IMSLex

2 Morphology

2.1 Maintaining the database

When building the morphological database, we started from the lexicon of a computational morphology system. In order to allow for a user-friendly extension of the database, we have developed a sophisticated acquisition tool for interactive lexicon maintenance (LexiTool, cf. [Lezius et al. 1999]). When entering a new word, the user is asked the minimal number of questions necessary to infer the new word's inflection type. The questions are internally modelled as a decision tree. In most cases only the lemma has to be typed in, and questions are answered by marking the correct one among the suggested alternatives. To allow the user to check the entry for correctness, the resulting paradigm can be displayed by LexiTool. Currently, the lexicon comprises 65.000 entries and is extended continuously.

2.2 The database

The morphological part of the database contains the lemmas of the different word classes, the respective inflection classes, and some administrative information (such as date of entry). The structure of the basic table is illustrated in figure 2.

Each word class is stored in a table derived from this basic table plus additional information, e.g. participle construction for verbs. As a consequence, different word classes can be combined in a single SQL query. In the following example, a join is performed on two tables, showing all the nouns ocuuring in the database as a common noun and proper noun.[1]

```
SELECT Substantive.Lemma FROM  Substantive, Eigennamen
                         WHERE Substantive.Lemma = Eigennamen.Lemma;
```

field	type	length	description
Lemma	text	variable	
Oberflächenform	text	variable	suppletive form: *ging* → *geh(en)*
Flexionsklasse	text	variable	inflection type according to morphology system
Name_Eintragender	text	variable	user name; short form
Eintragsdatum	date	fixed	date of entry; e.g. 10/15/1999

Figure 2: Structure of the basic table of the morphological section of IMSLex

To illustrate how linguists can profit from the resource in combination with the expressiveness of SQL, the following example shows how to determine all verb stems in the resource that can be concatenated with the suffix *-bar* resulting in an adjective which is also contained in the database (*wunder(n)* → *wunderbar*; engl. *(to) marvel* → *marvellous*).[2]

```
SELECT Lemma FROM  Verben
             WHERE Lemma IN
                        (SELECT rtrim(rtrim(rtrim(Lemma,'r'), 'a'), 'b')
                         FROM Adjektive WHERE Lemma ~ '.+bar$');
```

2.3 Querying the database

To make querying the database as easy as possible, we have developed a user-friendly tool (LexiExplorer, cf. figure 3) which combines an SQL-based and a query-by-example-based visual interface. Thus no specific database knowledge (and no knowledge of SQL) is necessary for posing simple queries.

3 Syntax

3.1 Extraction of subcategorisation frames

Besides morphological information, the lexical database provides for a second type of information that is essential for syntactic applications such as parsing. This includes information about subcategorisation frames and different kinds of distributional information, such as nouns that can be used in measure expressions (e.g. *Liter* in *1 Liter Wasser* (1l of water)), special adverbs (e.g. *rund* in *rund 100 Leute* (about 100 people)), adjectives usable only attributively or only predicatively, etc. Among these, subcategorisation frames for verbs are most prominent: each canonical sentence contains at least one verb and furthermore, subcategorisation frames are difficult to predict because of their broad variety (the IMS lexical database currently lists about 350 different types of frames).

Subcategorisation lexicons supporting wide coverage grammars can be extracted semi-automatically from corpora (cf. [Eckle 1999]; the lexical database currently contains

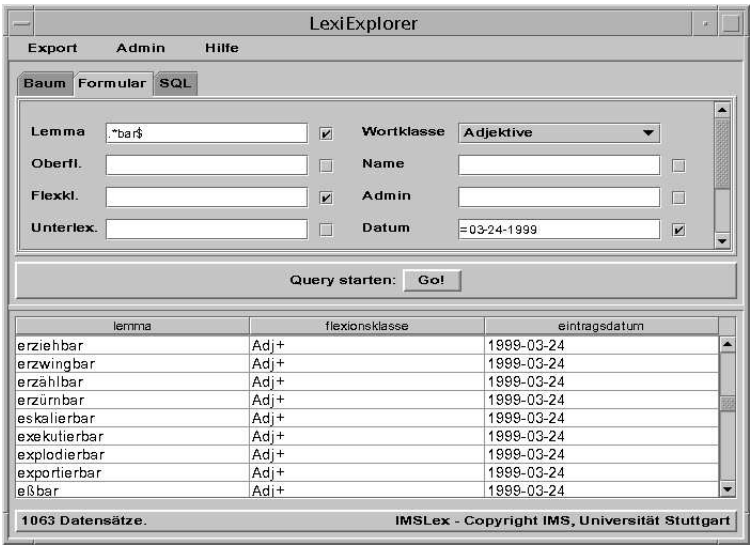

Figure 3: Database exploration assisted by the tool LexiExplorer

subcategorisation information for about 14.000 lemmas). However, these lexicons are necessarily incomplete since they only contain frames that showed up in a corpus. In the following section we present semi-automatic approaches for the completion of the lexicon.

3.2 Rules and heuristics for completion of the lexicon

The architecture we present provides for operations on the set of all subcategorisation frames of a lemma.[3] These operations allow for systematic adding of missing frames. An example is provided by verbs with locative complements: suppose that for a given verb, in a corpus you find a significant number of frame instances that only differ with respect to the preposition that is subcategorised. If furthermore these prepositions are locative ones, one can hypothesise that the verb subcategorises for a locative complement. Thus all missing locative prepositions (i.e., the ones that did not show up in the corpus) can be added to the lexicon.

In the same way, several frames may be collapsed into one in case the distinction between two or more frames is not relevant for the application in question; e.g. frames that are identical modulo a correlative *es* can be collapsed for parsing, following [Berman et al. 1998]. Likewise the mechanism can be used to set preference marks. An example are transitive verbs whose objects can be omitted. In case a sentence with such a transitive verb is ambiguous and allows for a reading as both transitive and intransitive, the transitive reading can be marked as preferred. Thus in a sentence like *Heute ißt Hans Brot* (today, Hans eats bread), *Brot* will preferably be analysed as object (and not e.g. as the last name of *Hans*). Note that only the transitive frames of verbs with an alternative intransitive reading will be marked as preferred.

3.3 Mapping to LFG representation

The subcategorisation frames contained in the lexical database are encoded in the TSNLP format (cf. [ESTIVAL ET AL. 1995]). At the IMS, we support parsing in the framework of LFG. For this application, the frames are mapped to a representation processible by the LFG grammar. Since frames are often added or modified, this mapping is done automatically and in a modular way. Thus it is guaranteed that after any modification of the lexical database, the LFG lexicon can be updated immediately and hence can take advantage of any improvement of the lexical database.

The frames are mapped both to macros that supply the verb's predicate-argument structure in LFG format as well as to macros providing for other information relevant for parsing, such as case restriction on nominal arguments, restrictions on sentential and prepositional arguments, etc. In our application, diatheses like passive are also generated during the conversion. The mapping from the TSNLP format to the LFG format of the example verb *essen* (to eat) is illustrated in the figures 4 and 5.[4]

Lemma	Subcategorisation frame	Example
essen	(subj(NP_nom),obj(NP_acc))	*Hans ißt Brot. (Hans eats bread.)*
essen	(subj(NP_nom))	*Hans ißt. (Hans is eating.)*

Figure 4: TSNLP-format of the subcategorisation frames of the verb *essen*

Lemma	Subcategorisation frame	Example
essen	{ @(NPnom-NPacc essen)	*Hans ißt Brot. (Hans eats bread.)*
	| @(null-NPnom-PASSIVE essen)	*Brot wird gegessen. (Bread is eaten.)*
	| @(NPnom essen) @(DISPR)	*Hans ißt. (Hans is eating.)*
	| @(null-PASSIVE essen) @(DISPR)	*Heute wird gegessen.*
	}	*(Today, one eats.)*

Figure 5: LFG-mapping of the subcategorisation frames of the verb *essen*

4 Using the resources in parsing

For applications like parsing, not only subcategorisation information is necessary but morphological information as well. Both are provided by the lexical database.

When parsing a sentence, each word first is analysed by a morphological component (derived from the lexical database). The morphological analysis provides information about the word's lemma form, its part of speech, and inflection. In a second step, the subcategorisation lexicon is looked up for information about the lemma. Finally, the LFG grammar analyses the sentence, taking into account both morphological and subcategorisation information. In figure 6 an example with the verb *ißt* (eats) is depicted.

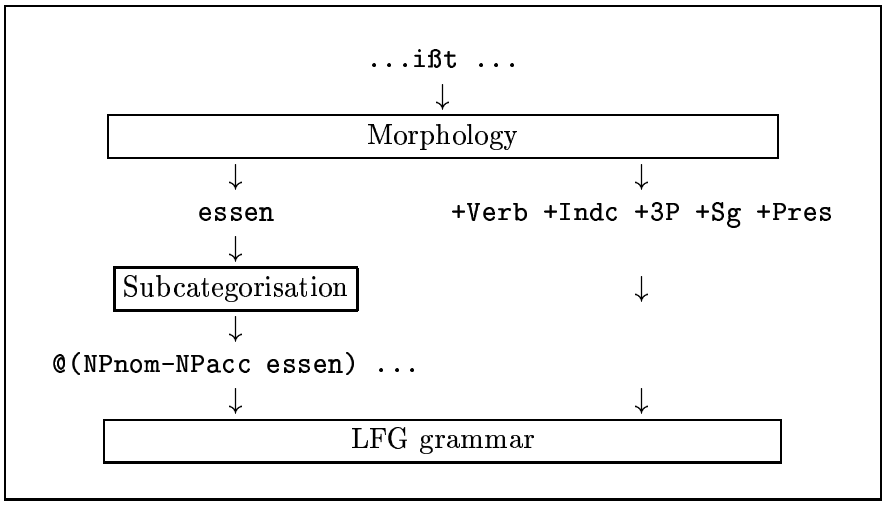

Figure 6: LFG analysis of the word form *ißt*

In our scenario, both the morphological component as well as the subcategorisation lexicon are derived from a common source, the lexical database. The main advantages are:

- The lexicon used by the morphological component and the subcategorisation lexicon have to be synchronised with respect to the lemma forms. That is, the lemma forms used in both lexicons must be identical (e.g. *essen* in figure 6). This requirement is easier to fulfil if both lexicons start out from one common source.
- The LFG grammar developed at the IMS is used for parsing newspaper texts. Typically these texts contain many neologisms, so a well organized lexical database is extremely important. This guarantees that (i) new entries can be easily added to the database, (ii) the entries are consistent, and (iii) applications like parsing can immediately take advantage of all improvements, on the level of both morphology as well as subcategorisation.

Conclusions

We have presented an architecture for a lexical database comprising morphological and syntactical information. The next step is to integrate semantic information.

Notes

[1]Literally: "Retrieve all lemmas from both tables, that of nouns (`Substantive`) and that of proper names (`Eigennamen`), where the lemma form of the noun is identical to the lemma form of the proper name."

[2]Literally: "Retrieve all lemmas from the verb table, where the lemma is enclosed in the following list: all lemmas from the adjective table which end on *-bar*, shortened by the suffix *-bar*."

[3]This property distinguishes the operations described here from traditional lexical rules like rules for passivisation. These rules only apply to *one* subcategorisation frame of a verb at a time, as opposed to a *set* of frames.

[4]TSNLP format: for each complement, its function (e.g. *subj*) and its category (e.g. *NP_nom*) is specified. Complements are separated by commas. LFG format: the notation is based on XLE, the grammar development environment of Xerox, which is used in the LFG parsing project (ParGram: `http://www.parc.xerox.com/istl/groups/nltt/pargram/`). `@(···)` represents a macro call, { ... | ... } indicates a disjunction.

References

[Berman et al. 1998] Judith Berman, Stefanie Dipper, Christian Fortmann, Jonas Kuhn (1998). "Argument Clauses and Correlative 'es' in German: Deriving Discourse Properties in a Unification Analysis", in: M. Butt and T. H. King (eds.) *Proceedings of the LFG98 Conference*, CSLI Online Publications. URL: `http://csli-publications.stanford.edu/LFG3/`

[Bröker/Dipper 1999] Norbert Bröker, Stefanie Dipper (1999). "Zur Konstruktion von Lexika für die maschinelle syntaktische Analyse", in: Gippert, J., Olivier, P. (eds.) *Multilinguale Corpora – Codierung, Strukturierung, Analyse*, 11. Jahrestagung der Gesellschaft für Linguistische Daten-Verarbeitung. Enigma corporation, Prag.

[Eckle 1999] Judith Eckle (1999). *Linguistisches Wissen zur Lexikonakquisition aus deutschen Textcorpora* Ph.D. thesis, Universität Stuttgart.

[Eckle/Heid 1996] Judith Eckle, Ulrich Heid (1996). "Extracting raw material for a German subcategorization lexicon from newspaper text", in: *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX 1996.*

[Estival et al. 1995] Dominique Estival et al. (1995). *The Construction of Test Material* TSNLP Report (WP 3.1). URL: `http://cl-www.dfki.uni-sb.de/tsnlp/publications.html#wp3.1`

[Lezius et al. 1999] Wolfgang Lezius, Arne Fitschen, Ulrich Heid (1999). "Datenbankbasierte Verwaltung und Pflege Morphologischer Information im IMSLex", in: Gippert, J., Olivier, P. (eds.) *Multilinguale Corpora – Codierung, Strukturierung, Analyse*, 11. Jahrestagung der Gesellschaft für Linguistische DatenVerarbeitung. Enigma corporation, Prag.

[Lezius 1999] Wolfgang Lezius (1999). *Das IMSLex – Online-Informationen*
URL: `http://www.ims.uni-stuttgart.de/projekte/IMSLex/`

[Schiller 1996] Anne Schiller (1996). "Deutsche Flexions- und Kompositionsmorphologie mit PC-KIMMO", in: Hausser, R. (ed.) *Linguistische Verifikation – Dokumentation zur Ersten Morpholympics 1994*, S. 37-52, Niemeyer, Tübingen.

[Schmid 1994] Helmut Schmid (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees", in: *Proceedings of the International Conference on New Methods in Language Processing 1994.*