# *ELEXIKO* – A Lexical and Lexicological, Corpus – based Hypertext Information System at the *Institut für Deutsche Sprache*, Mannheim

**Dr Annette Klosa, Dr Ulrich Schnörch, Dr Petra Storjohann**
Project *elexiko*
Institut für Deutsche Sprache
R 5, 6-13
68161 Mannheim
Germany
klosa@ids-mannheim.de, schnoerch@ids-mannheim.de, storjohann@ids-mannheim.de

**Abstract**
*ELEXIKO* is a relatively new lexicological-lexicographic project based at the *Institut für Deutsche Sprache* (IDS) in Mannheim. The project compiles a reference work that explains and documents contemporary German; it was specifically designed for online publication (www.elexiko.de). The primary and exclusive basis for lexicographic interpretation is an extensive German corpus. If one refers to *elexiko* as an Internet dictionary, it is purely for practical reasons. *elexiko* is (far) more than a dictionary in its traditional sense, although, of course, it contains descriptions of the meaning and use of a lexeme just as any traditional dictionary. It is both, a hypertext dictionary and a lexical data information system.

## 1 Writing the dictionary

Filling *elexiko* in modules is (besides our corpus-based approach) one of the two main lexicographic methods for our dictionary. *elexiko* is compiled not following the alphabetical order, but by analysing the semantic, syntactic, or morphological features of the lexicon systematically in batches. Thus, a complete word class, an entire word family, or a semantic field can be described systematically and separately. Furthermore, modules are also defined according to levels of frequency and distribution of lexemes in the *elexiko*-corpus. Modularity also means that in *elexiko* dictionary entries which have been written in other lexicographic projects at the *Institut für Deutsche Sprache* (cf. 2.) become an integrated part of the project. Along with publishing the list of headwords (taken exclusively from the *elexiko*-corpus) on the Internet in 2003, our dictionary was filled with sense independent information for each headword generated automatically or semi-automatically from the underlying corpus. This concerns 300,000 single-word entries comprising details on spelling, spelling variation, and syllabication. In a second step, the first 250 headwords, which were defined as the

demonstration module *(Demonstrationswortschatz)*, have been fully lexicographically described. The *Demonstrationswortschatz* primarily explains lexemes forming a semantic field around the core headword *Mobilität* (i.e., "mobility") and lexemes that are morphologically derived from *mobil* (see Figure 1) (e.g., *hochmobil* [i.e., "highly mobile"], *immobilisieren* [i.e., "to immobilize"], and *Mobilitätszentrum* [i.e., "center for mobility"]).

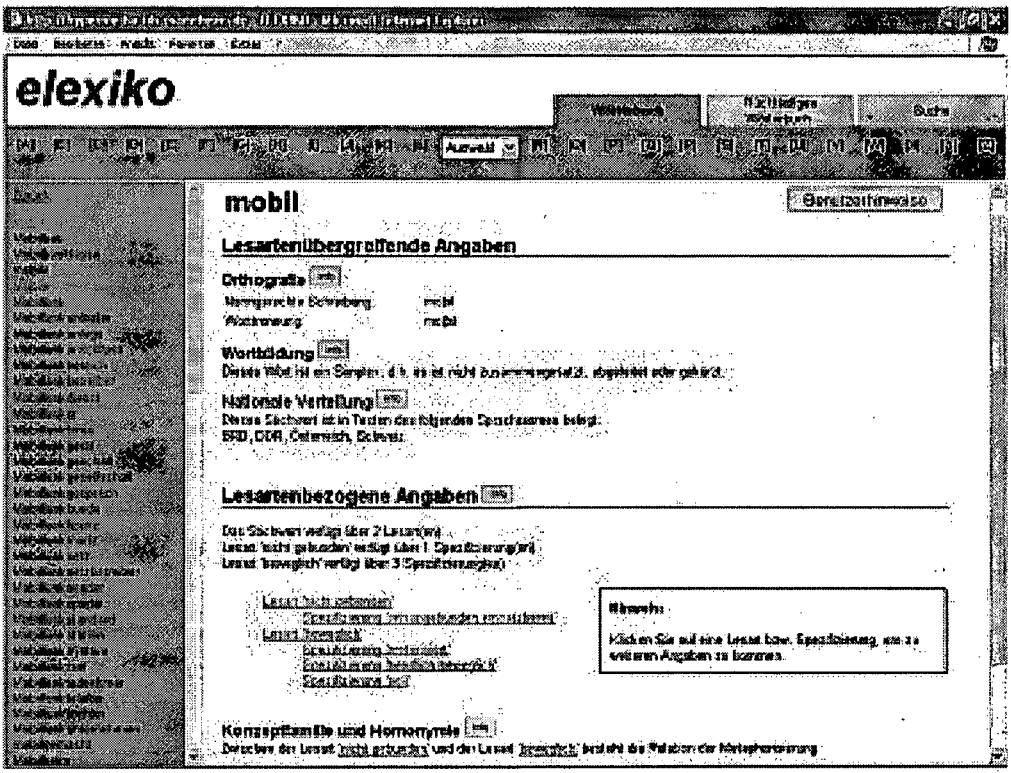


**Figure 1.** Sense-independent information for **mobil**

Moving on, the project began working on a module called *Lexikon zum öffentlichen Sprachgebrauch* (i.e., "dictionary on public discourse"). It contains aproximately 2,800 entries selected mainly by their (high) frequency in the *elexiko*-corpus, such as *Regierung, Arbeitsmarkt, Reform.*

Currently (March 2006), the dictionary contains approximately 500 fully lexicographically described entries. These entail sense-independent information on morphology and word formation *(Lesartenübergreifende Angaben)* as well as number of senses and their relationship (see **mobil** in Figure 1). They also offer a large scope of sense-related information *(Lesartenbezogene Angaben),* in detail: meaning definition, collocations, syntagmatic patterns, sense-related terms, pragmatics, and grammar (see **mobil** in Figure 2). With this wide spectrum of lexical information, *elexiko* exceeds other existing German dictionaries.

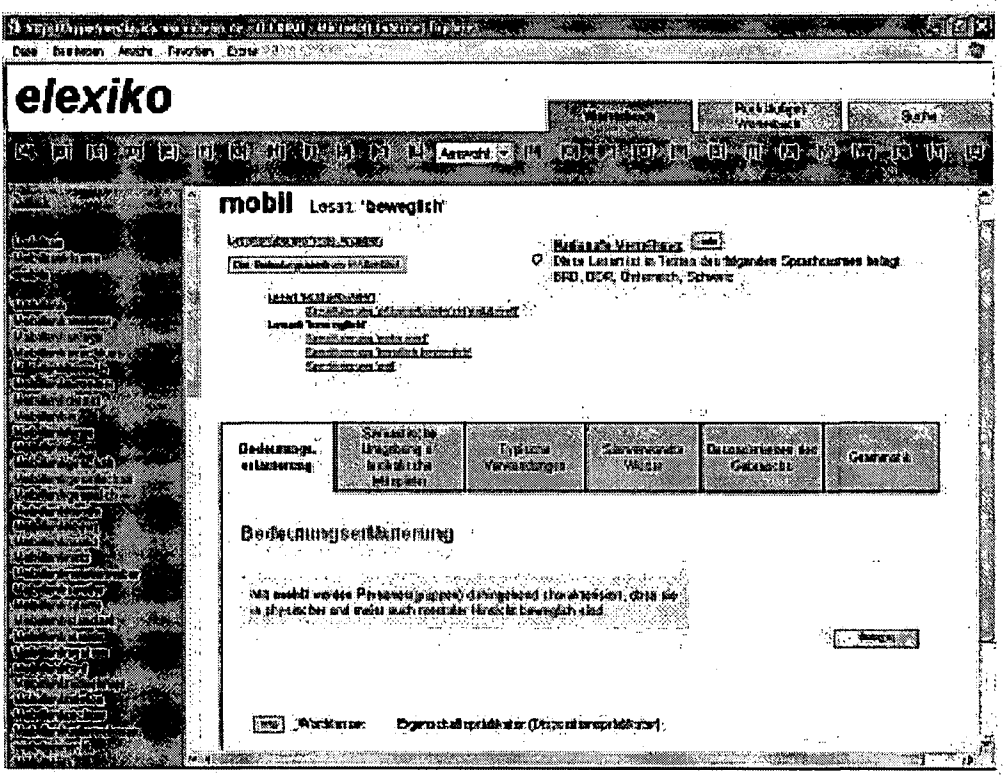


**Figure 2.** Sense dependent information for **mobil**

## 2 Future prospects

*elexiko* is growing continuously. In 2006, new modules will be incorporated: These are descriptions of over one hundred multi-word items (*Usuelle Wortverbindungen,* cf. http://www.ids-mannheim.de/lexik/UsuelleWortverbindungen/) connected with single-word entries of the *Demonstrationswortschatz,* approximately 700 German neologisms (*Neologismen,* cf. http://www.ids-mannheim.de/lexik/Neologie/) of the 1990s, and lexemes around the notional area designated by *guilt* (*Schulddiskurs,* cf. http://www.ids-mannheim.de/lexik/Zeitreflexion/) in Germany's postwar era 1945-1955. These modules either explain a specific part of the German lexicon which is not part of *elexiko* or focus on specific questions which are not covered by *elexiko*. At the same time, for a large number of low-frequent lexemes, information (e.g. on word formation) and corpus samples are generated automatically or semi-automatically by using various computer tools.

## 3 Technical background

For the process of writing and presenting *elexiko* on the Internet, we use numerous technologies and software tools. For example, the corpus query and processing tool COSMAS II (cf. http://www/ids-mannheim.de/cosmas2/) and its incoporated collocation programme

"Statistische Kollokationsanalyse und Clustering" are of particular benefit for numerous corpus-guided investigations within our practical working procedure. We also use ORACLE 9.i as a content-oriented database, and XSLT style sheets to generate an online presentation of entries. All lexical entries are structured XML-instances following a highly granular lexicographic data model (Document Type Definition) with over 400 tagging elements. This DTD has been developed specifically for the intended microstructure of our dictionary.

Our technical background enables us to offer the list of headwords and fully lexicographically described headwords on www.elexiko.de along with specific search features to users. Dictionary consultants not only find single headwords but can also look up groups of lexemes with the same semantic, syntactic, or morphological characteristics (see Figure 3, where the search-site of *elexiko* is shown with an inquiry concerning composed adjectives; the search-site will be extended continually).

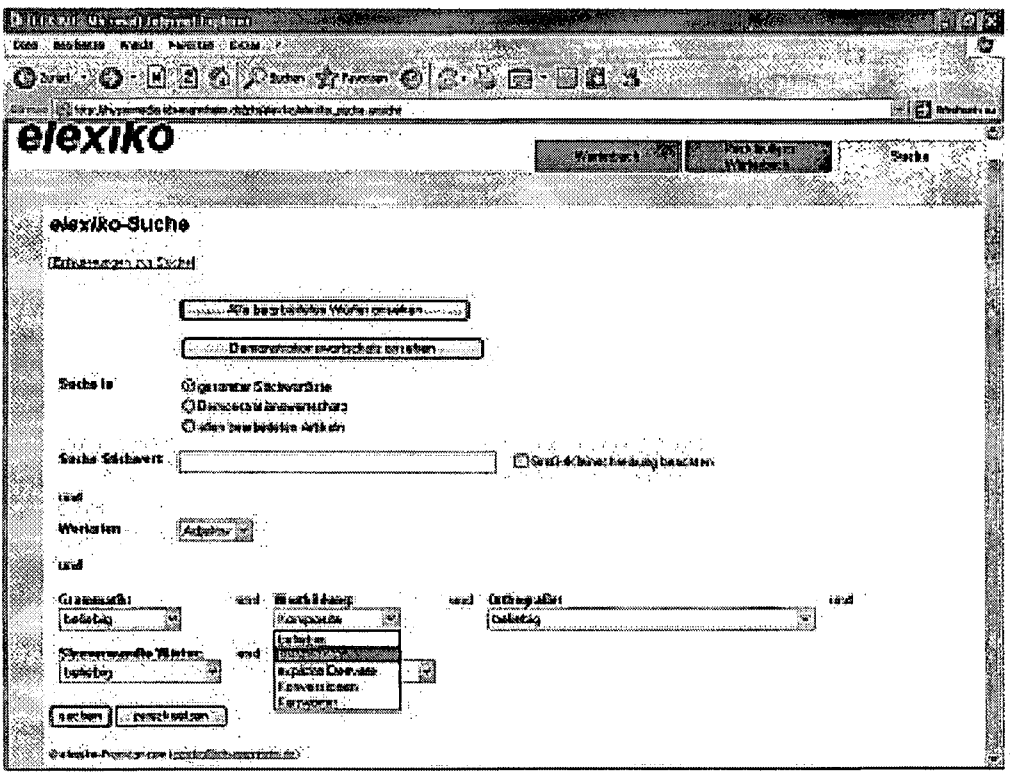


**Figure 3.** Search site of *elexiko*

**References**

**A. Dictionaries**

http://www.elexiko.de

**B. Other literature**

Belica, Cyril (1995), *Statistische Kollokationsanalyse und Clustering*. COSMAS-Korpusanalysemodul. Mannheim, IDS.

Haß, U. (2005), *Grundfragen der elektronischen Lexikographie. elexiko – Das Online-Informationssystem zum deutschen Wortschatz,* Berlin / New York, de Gruyter.

Klosa, A. (2005), '*elexiko*. Ein Onlinewörter buch zum Gegenwartsdeutschen', *Sprachreport* 3, pp. 6-9.

Storjohann, P. (2005), '*elexiko*: A Corpus-Based Monolingual Dictionary', *Hermes Journal of Linguistics* 24, pp. 55-82.