*Petra Storjohann**

elexiko: A Corpus-Based Monolingual German Dictionary

Abstract

This article provides an introduction to *elexiko*, the first German hypertext dictionary to be compiled on a corpus basis, which is currently being developed at the *Institut für Deutsche Sprache Mannheim* (IDS). First, a brief account of the design is given, followed by a demonstration of the methods and tools that are being employed to compile it. *elexiko* will provide not only an improved quantity of lexical information, but also a new quality of information which will be explained and illustrated at different levels of the microstructure of the dictionary. The description of word meaning and use in *elexiko* will be presented in detail, with a particular focus on the treatment of collocations, ambiguity, vagueness, and the presentation of senses. The development of a theoretically grounded procedure for lexicographic disambiguation is also described. This is then followed by a brief account of the treatment of grammatical details. Finally, issues of usability, the progress of the project and its future perspectives will be considered.

1. Introduction

The project *elexiko*, based at the *Institut für Deutsche Sprache Mannheim* (IDS), is currently developing a new dictionary of German and its present-day usage (cf. Haß-Zumkehr 2004).[1] The aim of the project is to explain and document contemporary German over the last 50 years, approximately the time span of two generations of speakers.

The research team combines corpus-based approaches with traditional lexicographic procedures and an awareness of current linguistic theo-

[1] For a brief description of the project see also: http://www.elexiko.de.

* *Petra Storjohann*
Institut für Deutsche Sprache Mannheim
R5, 6-13
D-68161 Mannheim
storjohann@ids-mannheim.de

ries. The language data used for lexicographic interpretation is retrieved exclusively from an extensive German corpus and presented in a hypertext structure which will be available publicly via the Internet. As well as being an 'electronic dictionary' through which a diverse spectrum of users can explore the German lexicon, *elexiko* will also be a linguistic data resource where linguists can extract data necessary for their research. However, *elexiko* must not be compared simply with digitised, machine-readable versions of printed dictionaries. Rather its hypertextuality and its underlying content-oriented lexicographic database facilitate a break with existing lexicographic conventions and open up a number of other ways of presenting data.[2]

With the help of a hypertext structure we will create an extensive linking system for illustrating various types of language structures. Semantically, morphologically or syntactically related lexemes will be linked up, pointing the reader to the main entry, sense or sub-sense of the related item. Hence, conventional signpost cross-referencing becomes redundant. Similarly, clicking on any paradigmatic or syntagmatic partner, on morphologically derived forms or on idiomatic phrasal expressions of a lexeme will take the user straight to the corresponding entry or to the corresponding occurrence within the entry.

In the underlying structure, each piece of lexicographic information is individually marked up (tagged). This has the benefit that specific linguistic details can be searched for selectively via a search and navigation system. At the same time, the systematic search for various types of information will realise a combination of semasiological and onomasiological presentations. Apart from conventional information such as meaning definition, syntax, examples of usage etc, the reader will always be merely a click away from answers to such questions as: In which contexts is this word used? What are typical dative complements of this verb? How many senses does this word have and how do they relate? Is this word bound to a specific type of text? Which words have negative semantic prosodies? Which idiomatic phrases describe the notion of 'dying'? How many words contain the suffix *–tät*? What are

---

2 For more details on the advantages of hypertextuality in lexicography see Haß-Zumkehr 2001, and Storrer 1998 and 2001.

typical compounds with *Haus*? Which words are specifically Austrian lexemes? How many words have alternating gender forms and more than one plural?

These questions are just illustrative examples to show what kind of linguistic information the user will be able to extract. A further advantage of using a hypertext structure is the possibility of embedding film, sound and picture documents (e.g. for illustration of meaning, function or for audible pronunciation). Furthermore, within a hypertextually structured dictionary any changes (e.g. newly conventionalised senses or contextual realisations) can be easily implemented without waiting for the next edition to be published. As such, continuous changes are possible which will result in accurately revised entries for new versions of the dictionary.

Overall, these benefits presuppose a lexicographic practice which rests on a number of key elements: an extensive corpus serving as an empirical basis; computerised tools that assist the search of the corpus; the modelling of a complex Document Type Definition; a complex data base that stores dictionary entries in a mark-up language (i.e. as XML-documents); and the development of a comprehensive navigation and search system.

2. Foundation and Methods

2.1. Corpus

While in English lexicography the compilation of a dictionary based on electronic corpora has a longer tradition (e.g. Cobuild project), in German lexicography corpus-driven approaches are a fairly recent development. *elexiko* will compile its data exclusively from a monitor corpus which was constructed for this purpose. The IDS Mannheim currently holds the largest collection of German corpora with a total volume of about 2000 million words.[3] It was using this foundation that an *elexiko*-corpus was built, which currently comprises about 1300 million words and is entirely based on written German. This volume

[3] A complete list of the German corpora at the IDS Mannheim can be obtained from http://www.ids-mannheim.de/kt/projekte/korpora/.

represents the most comprehensive data used to date for German lexicographic purposes. The corpus will grow and will be updated continuously thereby adapting to many different types of analysis and capturing instances of language change.

As far as the lexicographic process of describing lexemes and their uses is concerned, the corpus in *elexiko* is being used exploratorily.[4] Instances of natural language are studied in order to identify rules and patterns, and linguistic proto-typicalities which are then interpreted, evaluated and classified. Finding copious illustrative text samples is only a by-product of corpus-guided analysis. *elexiko* will monitor and document written German language use as depicted in the corpus. Some contrastive examinations have shown that corpus evidence often deviates from the information given in other dictionaries. These deviations consist primarily in a greater variety of semantic and syntactic patterns and in the non-normative alternatives which manifest themselves in actual language use. As linguists of a descriptive dictionary, it is our objective to supply descriptions of the diversity and complexity of lexical phenomena as they are evident in the corpus. We will account for different lexical forms and provide information on standard usage, as well as standardised variation and non-standardised occurrences in the corpus. Each item of information is accompanied by lexicographic interpretations, such as the proportion of individual variations that are standardised uses and commentary on forms of non-standardised use (e.g. old spelling conventions).

2.2. Corpus Processing

Words that are not very frequent in our corpus are easy to analyse without any further computer assistance. However, words that have a high frequency cannot be managed manually by any lexicographer. With the availability of mass data it is important to systematise that data, to arrange extracted results according to criteria in order to be able to di-

[4] As Sinclair (1991: 36) emphasises: ‘it is the possibility of new approaches, new kinds of evidence, and new kinds of description. Here, the objectivity and surface validity of computer techniques become an asset rather than a liability. Without relinquishing our intuition, of course, we try to find explanations that fit the evidence, rather than adjusting the evidence to fit the pre-set explanation’.

stinguish significant from insignificant data and to separate typical structures from atypical patterns. Computer technology facilitates this exploration of mass data and the compilation of an empirical foundation, from which information of a new quantitative and qualitative kind can be extracted. In our project, this role is performed by the corpus processing tool COSMAS, which was developed at the IDS and which works on the basis of mathematical statistical methods.[5] Apart from being a flexible search system, it is also a complex analytical tool yielding results which constitute indicative and measurable evidence which is then interpreted by the lexicographer who decides how to select irrelevant from relevant information. Although linguistic evidence is supplied by the corpus, most dictionary information has to be produced manually.

Of particular benefit in this regard is the concordancing software package *Statistische Kollokationsanalyse und Clustering,* developed at the IDS, which efficiently systematises data.[6] It performs large empirical explorations of data and detects linguistic structures by calculating the degree of lexical cohesion in the semantic and syntactic neighbourhoods of a word. It is possible to sort collocations and concordances according to different principles such as the degree of lexical cohesion, chronologically or by text source. The result is a retrieved list of co-occurrences which provide systematic access to corpus evidence. However, in some cases the examination of a word needs to go beyond the analysis of concordances. In such cases, the lexicographer needs to consult larger stretches of context. Using a collocation analysis and consulting concordances for lexicographic purposes significantly improves, simplifies and systematises the lexicographer's work.

---

5 COSMAS (Corpus Search, Management and Analysis System) was developed at the IDS by Cyril Belica (1995-2000) and is publicly accessible via the Internet: http://www.ids-mannheim.de/cosmas2/.

6 See: http://www.ids-mannheim.de/kt/projekte/methoden/ka.html.

3. Compiling the dictionary

3.1. Headword List

Headwords for this dictionary were compiled on the basis of the underlying *elexiko*-corpus. For this purpose, lemmatising software was employed, which sorts all occurring word forms and grouped them according to the lexeme from which they were derived.[7] Through this method we were able to document all existing words that occur in our corpus. The main criterion for including the retrieved headwords in *elexiko* was a specific corpus frequency which guarantees a sufficient number of records for the lexicographic analysis of each word.

As the list generated proved only partially accurate and showed inadequate lemmatisation in a number of cases, all headwords were manually analysed. In this manner, non-German words, typographical errors, initials, numerals and other forms falsely lemmatised by the software were manually detected and, if necessary, deleted from the list or included in the list after correction. Since the German spelling system recently introduced new spelling regulations, spellcheck software was also applied. The conversion from old to new spelling was conducted with the help of two software programmes (*Duden Korrektor Plus* and *Corrigo*)[8], the output also undergoing an additional editorial process. This procedure enabled us to proceed with a list of around 300,000 potential headwords for *elexiko* which will be further expanded at a later stage by the inclusion of other lexical items such as multi-word units and morphological elements. With this number of headwords the dictionary surpasses Duden-GWDS (1999) which contains around 200,000 headwords.

3.2. Principle of Modularity

elexiko will be compiled by adopting the principle of modularity, meaning that the lexicon will be analysed systematically in batches, so-called modules, which are defined by specific semantic, syntactic or

---

7 See: http://www.ids-mannheim.de/kt/projekte/methoden/gl.html.

8 For details on *Corrigo* see: http://www.clt-st.de/produkte/corrigo.html.; for information on *Duden Korrektor Plus* see: http://www.duden.de/index2.html?produkte/elektronisch/korrektor/korrektor.html.

morphological criteria. The principle of modularity has a number of advantages. Specific lexicographic analyses can be distributed to linguists with specific expertise. Furthermore, a modular compilation can systematically and consistently describe words that are interrelated in any way and the result be presented simultaneously. Another great benefit of this method of working is the immediate inclusion of cross-references and the systematic linking to related words in order to illustrate lexical structures within a specific part of the lexicon.

Using this approach, semantic fields, entire word families, or a complete word class will be described systematically and separately. Other modules, however, might be defined for the inclusion of information for the entire lexicon, particularly information which can be extracted automatically for each headword.

In an initial step, the dictionary is filled horizontally, thus filling each entry with information generated automatically or semi-automatically from the corpus. This includes details on syllabication and spelling alternatives. A second module deals with the detailed description of the first 250 headwords which are defined as the demonstration module (*Demonstrationswortschatz*). The *Demonstrationswortschatz* contains lexemes which construct a semantic field with the core headword *Mobilität*. In addition, headwords were added according to various criteria. For example, lexemes that are morphologically related to *Mobilität* (e.g. *hochmobil, immobilisieren, mobilisierbar, Mobilitätsverhalten, Mobilitätszentrum*) and lexemes which are pragmatically stigmatised (e.g. *global, kreativ, modern, Reform, Stau*) were included. Furthermore, the number of lexemes from word classes which were disproportionately underrepresented was increased accordingly. The purpose of the *Demonstrationswortschatz* is to demonstrate the depth of semantic, pragmatic and syntactic information of an entry, to indicate the extent of the linking system and to offer initial insights into possible search options. At the same time, multi-word items containing some of the 250 words of the demonstration module are being compiled and described, and another module containing German neologisms of the 1990s is also being incorporated into *elexiko*.

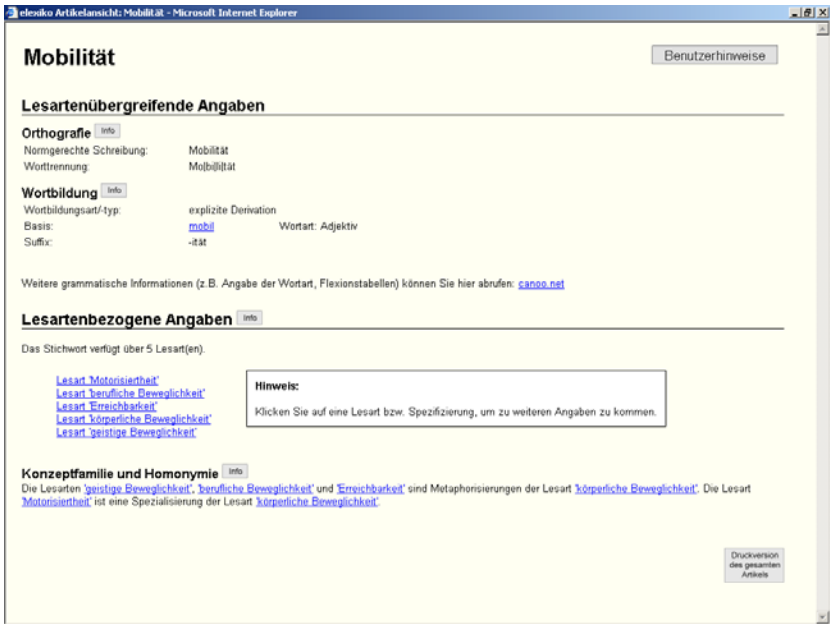
3.3. Microstructure

One chief task during the conceptual phase of this project was to develop a notion of the kind of lexical information to be included and the extent or depth of descriptions to be integrated. Essentially, each entry is divided into two main parts. First, sense-independent information of a lexeme is given, and secondly, information which is bound to a specific sense of the search item is provided. Examples that illustrate the use of a word in a specific way are taken exclusively from the constructed corpus and are attached to most important elements of an entry (see button labelled *Beleg(e)* on some following screenshots).

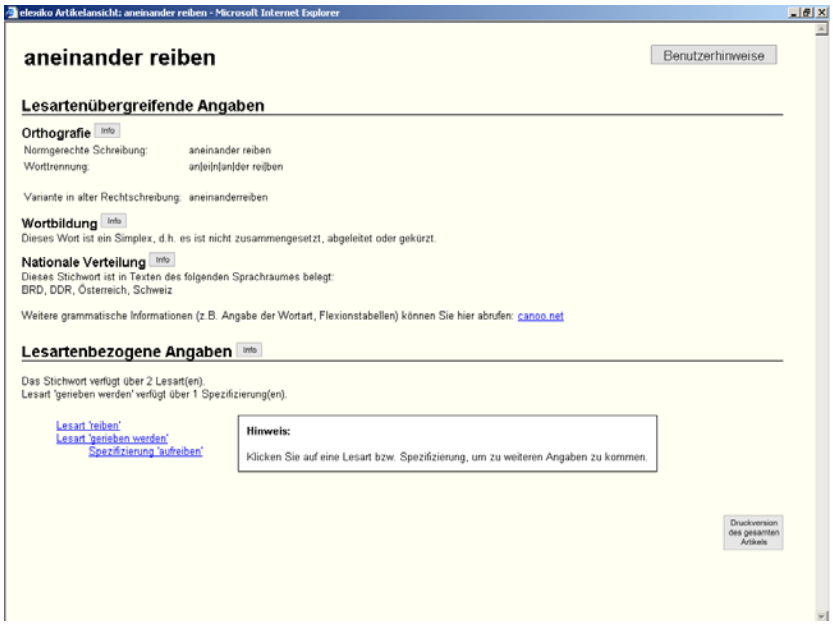
3.3.1. Sense Independent Information

Details that concern the lexeme itself will be presented sense-independently in the separate headword section. This focuses exclusively on information that applies to the entire entry and not to a specific sense and contains details on spelling, spelling variants and syllabication, morphological information, a link to *canoo*[9] for some grammatical details, diachronic information (not yet incorporated in the demonstration module), regional variation (if it applies for the lexemes themselves and not for a specific sense), a list of senses/sub-senses and their semantic interrelation. The following two illustrations (example 1 and 2), *Mobilität* and *aneinander reiben*, serve to demonstrate the presentation of sense-independent details in *elexiko*.

9 For further information on *canoo* see www.canoo.net.


elexiko Artikelansicht: Mobilität - Microsoft Internet Explorer
Mobilität
Benutzerhinweise
Lesartenübergreifende Angaben
Orthografie
Info
Normgerechte Schreibung:
Mobilität
Worttrennung:
Mo|bi|li|tät
Wortbildung
Info
Wortbildungsart/-typ:
explizite Derivation
Basis:
mobil
Wortart: Adjektiv
Suffix:
-ität
Weitere grammatische Informationen (z.B. Angabe der Wortart, Flexionstabellen) können Sie hier abrufen: canoo.net
Lesartenbezogene Angaben
Info
Das Stichwort verfügt über 5 Lesart(en).
Lesart 'Motorisiertheit'
Lesart 'berufliche Beweglichkeit'
Lesart 'Erreichbarkeit'
Lesart 'körperliche Beweglichkeit'
Lesart 'geistige Beweglichkeit'
Hinweis:
Klicken Sie auf eine Lesart bzw. Spezifizierung, um zu weiteren Angaben zu kommen.
Konzeptfamilie und Homonymie
Info
Die Lesarten 'geistige Beweglichkeit', 'berufliche Beweglichkeit' und 'Erreichbarkeit' sind Metaphorisierungen der Lesart 'körperliche Beweglichkeit'. Die Lesart 'Motorisiertheit' ist eine Spezialisierung der Lesart 'körperliche Beweglichkeit'.
Druckversion des gesamten Artikels

Example 1



Example 2

(1) Spelling and spelling variety

With *elexiko* covering around 300,000 words, it offers readers a comprehensive lexicographic volume, containing more words than any other contemporary German dictionary (*Duden-GWDS* contains about 200,000 head words). Although old spelling conventions are statistically more common, each word will appear in its new spelling.[10] At the same time spelling variants which occur in the corpus and are rule conforming (e.g. *geografisch* vs. *geographisch*, *potentiell* vs. *potenziell*, *Smalltalk* vs. *Small Talk*), misspellings, if they show a regular occurrence (e.g. *Aquisition* vs. *Akquisition*), and old spelling conventions (see example *aneinanderreiben* vs. *aneinander reiben*) are supplied.

[10] German underwent new spelling regulations to be statutory since August 1998. As the corpus covers language data since 1946, older spelling variations are statistically more common.

(2) Word formation and variant forms

Information on word formation is not only a good source to extend one's vocabulary but also an opportunity to illustrate a network of morphologically related words. In *elexiko*, the user will, for example, find the corresponding source of derivational forms. Variant forms of a word (*andernfalls vs. anderenfalls, mickrig vs. mickerig*) and morphological constituents of compounds and derivations are described systematically. At a later stage, information on the productivity of word formation (i.e. any morphologically related form of a lemma) will be included. Provided that corresponding related forms are present in the corpus, entire morphological families can be traced and documented, so that the entry for the adjective *mobil*, for example, lists any derived form (e.g. *immobil, immobilisieren, mobilisierbar, mobilisieren*) and compounds (e.g. *mobilbehindert, mobiltelefonisch*). In contrast, print dictionaries are restricted in their illustration of the extent of word families by the paucity of space. Each documented form itself will be a separate entry and further details of morphologically related words will then be obtained by choosing the link provided.[11]

(3) Diachrony

Although *elexiko* is a synchronic dictionary focussing on contemporary German, it will also provide diachronic information. The diachronic part is divided into three sections: a diachrony from Old High German to 1700; a diachrony from 1700 to 1945; and a so-called Microdiachrony covering the time span between 1946 to present day. Each part will contain different types of information. The first part (from Old High German to 1700) will mainly provide etymological information by referring to well-known reference works. For the diachrony between 1700 and 1945 some particular key words will be semantically described by consulting historical IDS-corpora.[12] For the description of lexemes in the microdiachronic part, contrastive analyses between historical and contemporary use will be conducted. As records can

[11] The outlined plan for the description of word formation is not yet fully incorporated in *elexiko* and will not be realised before summer 2005.

[12] See: http://www.ids-mannheim.de/lexik/HistorischesKorpus/.

be sorted chronologically with the help of COSMAS, we are able to look at decades individually in order to trace and document language change within a short period of time. This mainly concerns changes of a semantic, discursive, contextual or pragmatic nature, or changes in the frequency of occurrence. Our attention focuses on differences in the behaviour of the search item at a particular point in time and interpretations will mainly be presented in a narrative style.[13]

(4) Regional variety

Sense-independent information concerning regional varieties, by which we mean varieties of standard between Austrian and Swiss German compared to German as present in Germany, is also documented by the inclusion of an appropriate inventory of about 20% Austrian texts and 10% Swiss texts within the corpus. These proportions correspond to the respective percentages of people in German-speaking areas. Individual dialectal information, however, will not be supplied. An additional regional variety will be documented; namely the lexical characteristics of the language of former East Germany. As can be seen in the two examples, regional variety is either documented sense-independently (compare information 'Nationale Verteilung' in example *aneinander reiben*) or sense-dependently (see *Mobilität* in section 3.3.2).

(5) Conceptual Sense Relations

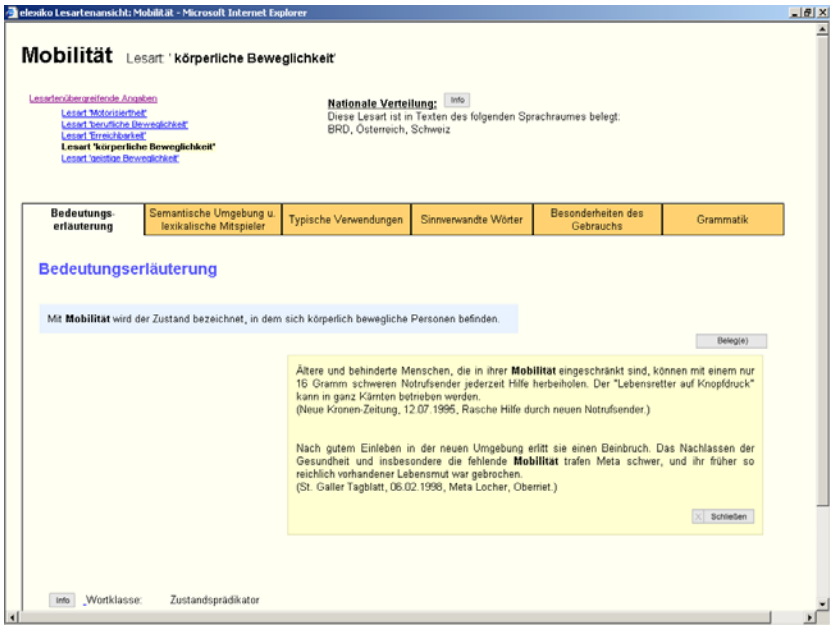
As the lexicon contains many ambiguous words, the interrelation between individual senses of such items will need to be determined. Different types of semantic relatedness between senses (e.g. metaphorisation, metonymy, irony, specification, generalisation etc.) will be elucidated to pinpoint conceptual families (see example *Mobilität* headline 'Konzeptfamilie und Homonymie').

3.3.2. Sense-Related Details

The main focus of sense-related lexicographic information will be on meaning, use and grammar. However, pronunciation (not yet incorporated), encyclopaedic details and regional restrictions can also be context-

[13] Information on diachrony is not yet incorporated into the *Demonstrationswortschatz* and will be added at a later stage.

dependent and are thus presented as sense-bound details. The following illustration (example 3) serves as a demonstration for sense-dependent details for the lexeme *Mobilität* in its sense of ‘physical mobility’ (‘körperliche Beweglichkeit’).



Example 3

(1) Meaning and Use

(1a) Presentation of Senses

In *elexiko*, words are accessed via their orthographic form; a distinction between homonymy and polysemy is not made; both phenomena are considered forms of lexical ambiguity. As a synchronic dictionary we cannot account for the etymologies of 300,000 lexemes, the origins of most of which remain obscure. Other criteria, such as semantic relatedness or a change in syntax have proven adequate for a clear cut of senses of some polysemous or homonymic words. However, they cannot be applied as formal criteria for the entire lexicon, as there are numerous cases where they fail to account for a distinction. Words, whether they

are polysemous or homonymic, are thus considered ambiguous and treated as single entries.

Dictionaries not only show little agreement as to the question of how many senses and sub-senses a lexeme has, but also use different criteria to arrange word senses according to a sense enumeration system (Ravin and Leacock 2000: 1).[14] However, such a presentation proposes a sense hierarchy, a kind of finiteness and completeness of senses. In *elexiko*, senses are not arranged according to an enumeration system, as such a system is inadequate in its demonstration of semantic overlap between senses or contextual sense specifications, as they define their senses in an atomistic way. As can be seen from the illustration above, senses or sense specifications (sub-senses) of a word will be offered in the form of guide words which signal to the user the conceptual-referential domain of each sense.

e.g. *Mobilität*:

Lesart: 'Motorisiertheit'
Lesart: 'berufliche Beweglichkeit'
Lesart: 'Erreichbarkeit'
Lesart: 'körperliche Beweglichkeit'
Lesart: 'geistige Beweglichkeit'

Guide words have so far not been used in any German monolingual dictionary, but they function as 'signposts' in a number of English dictionaries, particularly in EFL-dictionaries such as the *Cambridge International Dictionary of English* (CIDE), *Longman Dictionary of Contemporary English* (LDCE) and *Oxford Advanced Learner's Dictionary* (OALD).

(1b) Operational criteria for word sense distinction

One of the most important and difficult tasks of a lexicographer is to resolve lexical ambiguity. Lexicographers disambiguate the senses of a word by comparing linguistic patterns such as paradigmatic and syntagmatic structures (Reichmann 1989: 111-114) and by using a good proportion of their own intuition. We believe that with the accessibility

[14] Senses are arranged hierarchically following different aspects such as: core vs. peripheral senses; most common vs. uncommon senses; according to corpus frequency or degree of prototypicality etc.

of comprehensive data and computational tools, lexical disambiguation needs to be addressed from a different perspective.[15] What German lexicography has lacked so far is a theoretical basis upon which the different use of a word can be identified. We follow Moon (1987: 90) that 'no single method or criterion suffices: words vary too much in kind'. In the conceptual phase of our project, we have developed a disambiguation theory that consists of a multi-dimensional linguistic model (Storjohann 2003). We believe that only the interaction of different criteria will enable the classification of different lexical patterns and thus sufficient disambiguation of an ambiguous word.

Our disambiguation model contains different linguistic classifications that are interconnected. It primarily focuses on the sentential-functional behaviour of lexical items and thus illustrates the propositional differences that words can carry semantically.[16] As the use of a word is closely connected with its function within a proposition, the sentential context of a search item receives most attention. Whereas autosemantic words are grouped into sentential-semantic classes, which mainly consist of predicators[17] (v. Polenz 1988: 159-167), function words are disambiguated by their different syntactic functions in a phrase or sentence (following Zifonun et al. 1997). When a word portrays different functions in different contexts, and hence exhibits different linguistic categories within the linguistic model, different word senses are differentiated as demonstrated below:

e.g. *Einschränkung*:

- action-denoting predicator (Handlungsprädikator): sense 'das Begrenzen/das Vermindern' (e.g. *die Einschränkung des Angebots*)

---

[15] Disambiguation is understood in terms of the lexicographic procedure of identifying and distinguishing the senses of a word for further semantic/syntactic description in a dictionary entry.

[16] As Pustejovsky (1995: 62) emphasises: 'by defining the functional behaviour of lexical items at different levels of representation we hope to arrive at a characterization of the lexicon as an active and integral component in the composition of sentence meanings'.

[17] In *elexiko* a predicate is defined in terms of Propositional Logic and Predicate Logic. It designates a property or a relation and can be ascribed to different objects (cf. Seiffert 1969: 23).

- state-denoting predicator (Zustandsprädikator): sense 'Bedingung' (e.g. *eine bestimmte Einschränkung auferlegen*)
- quality-denoting predicator (Eigenschaftsprädikator): sense 'Behinderung' (e.g. *mit einer körperlichen Einschränkung leben*)

e.g. *beranken*:

- action-denoting predicator (Handlungsprädikator): sense 'etwas anpflanzen' (e.g. *jemand berankt die Wand mit Efeu*)
- event-denoting predicator (Vorgangsprädikator): sense 'bewachsen' (e.g. *Efeu berankt die Wand*).

e.g. *während:*

- preposition (Präposition): sense 'im Verlauf von' (e.g. *während der Demonstration passierte nicht viel*)
- conjunction adversative (Konjunktion adversativ): sense 'wohingegen' (*während die einen lachten, weinten die anderen*)
- conjunction temporal (Konjunktion temporal): sense 'als' (*während sie verreist waren, […]*)

The advantage of describing a lexeme according to functional classes lies in the illustration of the connection between the semantic/syntactic form and the propositional potential of a communicative unit (Strauß 1989: 788-796). Furthermore, the disambiguation model contains other systems with elements of a different sense-discriminating nature: referential-denotational information; and semantic or syntactic specifications.[18] In this way, lexicographic sense distinction has been elevated from a procedure conducted by introspection to a model-based task which linguistically justifies meaning discrimination.

(1c) Lexical Vagueness

Difference of meaning depending on the situational context is a practical problem in lexicography. It is usually not easy to define clear criteria to distinguish between a sense with invariant semantic properties

[18] Specifications are understood as semantic properties which are identified by complements/adjuncts (e.g. aspectual features (aktionsarten) for process-denoting predicators). Others, like quality-denoting predicators, can be subcategorised according to their specifications into emphasising, classifying and modifying predicators. As far as function words are concerned, they often carry functional specifications. Conjunctions for example function as connectors of clauses with specifications such as 'conditional' or 'concessive'.

and sub-senses with contextually determined information (Cuyckens & Zawada 1997: XV; Brisard et al. 2001: 262). Crucially, the study of corpus data reveals that the information derived from a corpus is much more differentiated with respect to the lexical vagueness of some words. Senses can be further determined semantically in their specific contextual use. Therefore, a sense is considered a general undetermined instance, whereas in contexts the word acquires its specific semantic manifestation and is then able to exhibit specific sub-senses. Such a notion of determinacy and indeterminacy is found in various semantic approaches (cf. Cruse 1986; Pustejovsky 1995; Cuyckens & Zawada 1997).

We believe that the illustration of senses and contextual sense specifications provides the user with a wider understanding of context, meaning and use. Therefore, semantic indeterminacy will be explored in terms of specific contextualisation (called *Spezifizierung*), and if discursive patterns occur in the actual contexts, they will be documented as specific discursive sense variations. However, they will only be presented as contextually determined sub-senses when they show a certain degree of habituality in the corpus. If the corpus provides sufficient evidence of a contextual specification, this sense variation will receive a full semantic-pragmatic and syntactic description in *elexiko*.

e.g. *Arbeit*

Lesart:	‘Tätigkeit’
	Spezifizierung: ‘Erwerbstätigkeit’
	Spezifizierung: ‘Training’
Lesart:	‘Mühe’
Lesart:	‘Werk’
	Spezifizierung: ‘Kunstwerk’
	Spezifizierung: ‘Klassenarbeit’
Lesart:	‘Arbeitsplatz’

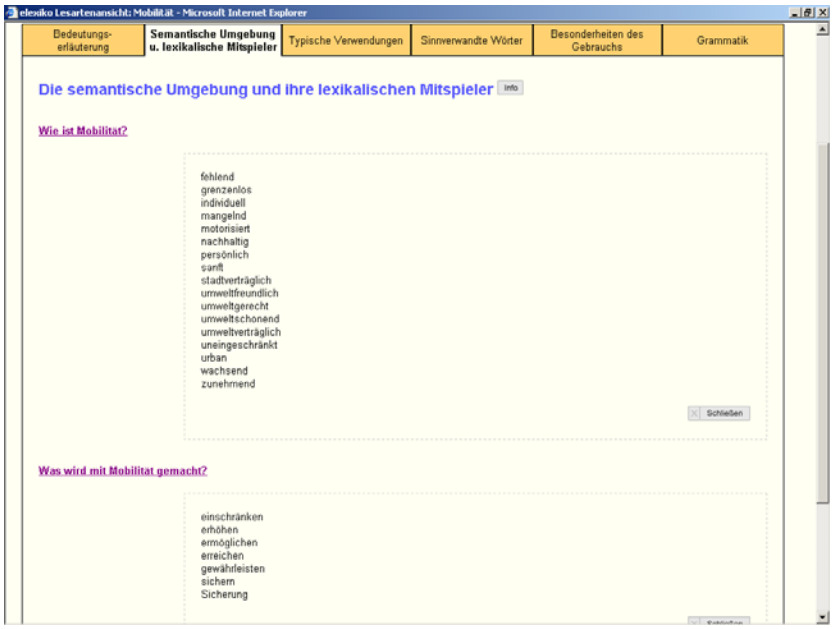
(1d) Collocations[19]

Collocations often represent habitual syntagmatic or paradigmatic patterns of the search item. The extraction and investigation of collocations

[19] Collocations are understood as the co-occurences (salient words) that co-occur with the search item in a specific context. In this paper, collocation is not associated with grammatical choices but defines lexical relationships.

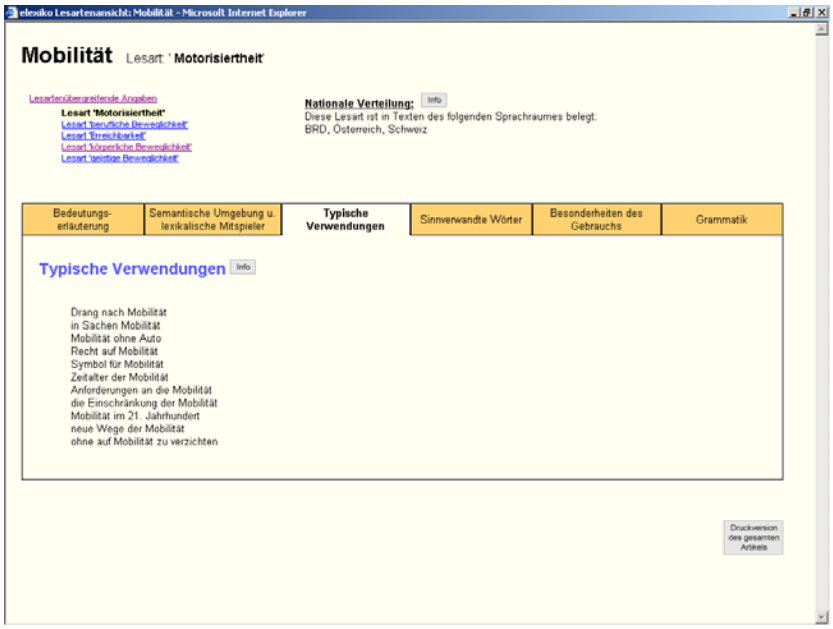
help us to identify the contextual environment and use of a word. As the use of a word is closely connected with its senses, the different syntagmatic and paradigmatic patterns are presented sense-bound. They are included to demonstrate semantic and syntactic variations, flexibility and/or constraints, information which is of interest to many language learners.

Syntagmatic patterns will be sorted primarily with respect to the semantic argument structure of a word. Typical syntagmatic lexical elements that accompany a word semantically in a context, and for example function as agents of verbs or modifications of nouns, are key information within the semantic description of a lexeme. With the help of question frames such as *Who does X? Who is affected by X? How is X modified? How is X typically characterised?* collocational partners are grouped according to syntactico-semantic functions and thematic fields. They must, however, not be compared to the traditional syntactic notion of an argument structures; our focus is entirely on the semantic company of a specific word. Especially for ambiguous lexemes, a sense-related presentation of the argument structure has the benefit that syntagmatic patterns of the different senses can be compared and differences in use can be more easily perceived. (For an overview see *Mobilität* sense ‘Motorisiertheit’ [‘with means of transportation’]).



Example 4

More complex syntagmatic structures which are usual, and thus statistically significant, patterns will be presented in their full syntagmatic phrasal form (see heading *Typische Verwendungen* in illustration). However, they will not include idiomatic phrases, as these are described separately. The focus will be on typical structures that demonstrate prepositional complements, typical adjuncts or any typical complements and supplements of verbs.



Example 5

In analogy to syntagmatic patterns, paradigmatic structures vary from sense to sense (or even from one contextual specification to another), as they are restricted to specific contexts. As non-native speakers in particular will benefit from a sense-bound presentation of sense relations, it is all the more surprising that, with the exception of including synonyms or antonyms in definitions, so far only a few German dictionaries have provided a sense-related presentation of paradigmatic relations. *elexiko* will also present a differentiated system of paradigmatic relations including various subtypes of incompatibility (cf. Cruse 1986, Lutzeier 1981) and vertical structures such as hyponymy and partonymy. Again, these are retrieved through the study of collocations and concordances. In some cases, however, paradigmatic investigations require the examination of a larger context. As sense relations are context-dependent, their presentation is always illustrated with accompanying examples from the corpus.

Example 6

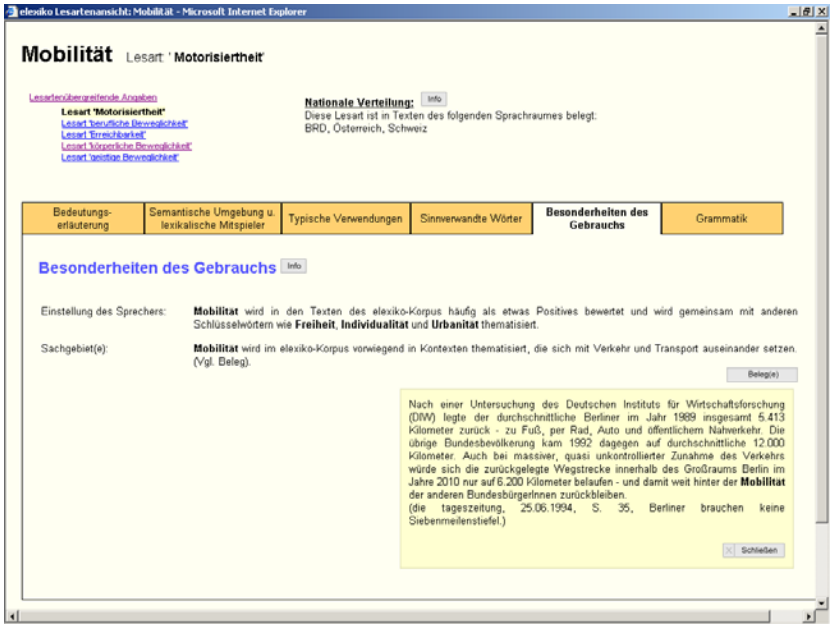
The presentation of lexical structures is one of our central objectives so that the reader acquires a better perception of the interrelatedness of words within the lexicon. The linking to other elements of the lexicon is hence a central aspect of our lexicographic enterprise. For instance, entire paradigmatic fields can be captured or all corresponding lexemes derived from one specific lexeme viewed. Synonyms, antonyms, hyponyms, etc, which themselves obtain a separate entry, are connected systematically. A complex system of cross-references, which are actual hypertext links facilitates the visual presentation of lexicon structures and provides direct access to related items. Links to paradigmatic partners are provided to the corresponding sense or sense specification of the related lexeme.[20] The lexeme *Mobilität* in the sense 'Motorisiertheit' has the complementary partner *Immobilität*.

[20] Links are provided to the related word entry instead of the sense, as long as the related paradigmatic partner has not been fully described lexicographically.

This paradigmatic relationship is restricted to a common verbal aspect (cf. Lutzeier 1981) and hence the link provided will lead the user to the entry *Immobilität* in its sense ‘Unmotorisiertheit’.

(2) Pragmatics

In *elexiko* the meaning, use and pragmatic force of a word are considered a close unit. Hence we follow a sense-related description of pragmatic features where pragmatic functions of a word are documented for a particular word sense. These mainly concern references to domain, contextual situations, type of text or constellation between speaker and hearer. Information on utterances which contain a specific intention of the speaker or have a certain pragmatic effect on the discourse is also included (see example 7).

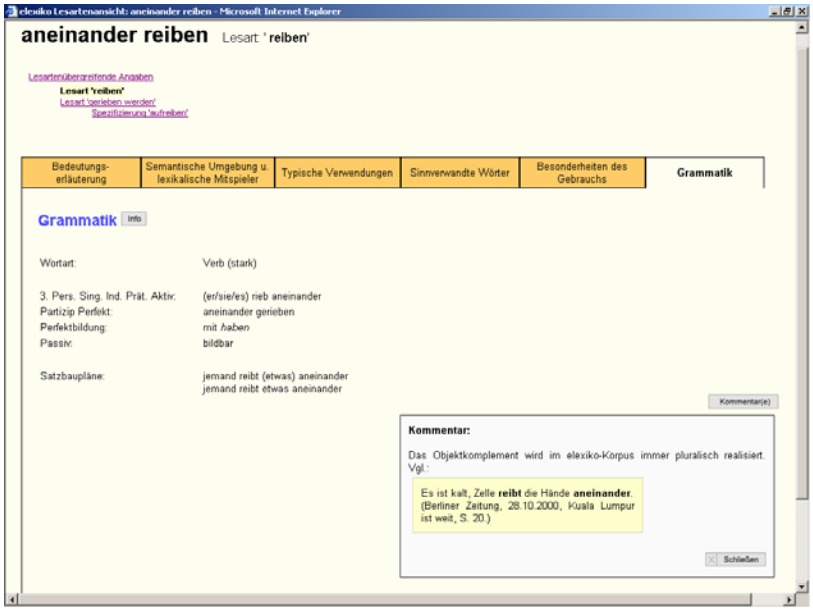


Example 7

(3) Grammar

The illustration of a word and its various multiple syntactic forms is very restricted within a print dictionary. Overall, the presentation of gram-

mar in German monolingual dictionaries does not show the deep interrelation between semantics and grammar, since grammatical information is often provided for the lexemes as a whole and not for individual senses of a word. *elexiko* is primarily concerned with the actual and current use of German which is considered to include its semantic and syntactic patterns and functions within a phrase or sentence. This is best illustrated by capturing syntactic structures in their corresponding semantic environment in order to illustrate necessary components that constitute the syntactical aspects of the use of the word. As semantics, grammar and lexical functions are considered intertwined linguistic components of a lexeme, they are shown in a sense-related presentation (cf. Pustejovsky 1995).



Example 8

Grammatical categories, such as gender, typical verb patterns, gradability of adjectives etc, occur in most monolingual dictionaries and will also be included. In addition, information on grammatical variation (e.g. alternating gender or plurals, syntactic phrase structures) and its

relative corpus frequency is given to provide an insight into the proportionate relationship between different variations.

Summary

The model entries shown as screenshot illustrations in this paper serve only as examples, all of which have their own specific linguistic details. Other entries, for example, include specific encyclopaedic information, information on other reference works that described the same lexeme or some entries include photographs. In other entries a number of lexicographic explanations and comments are attached to individual semantic or grammatical information. Diachronic details or complex details on word formation, as outlined in 3.3, are not yet included in the entries of the demonstration module. In addition, sound or video documents are not integrated into this module yet.

The corpus-driven approach offers a vast amount of material for linguistic analysis. As a hypertext dictionary *elexiko* is not confronted with the space problems of print dictionaries. Hence, a lexeme can be described in more detail. Overall, *elexiko's* entries contain a depth of information that cannot be found in any other electronic German monolingual dictionary (compare the entry *Mobilität* with the same entry in the electronic version of *Duden-GWDS* [2000]):

> **Mo|bi|li|tät,** die; - [lat. mobilitas, zu: mobilis, mobil]: **1.** (bildungsspr.) *[geistige] Beweglichkeit:* Die M. der Vierziger ist drastisch eingeschränkt (Schreiber, Krise 37); seine Argumentationen zeugten von hoher M. **2.** (Soziol.) *Beweglichkeit (in Bezug auf den Beruf, die soziale Stellung, den Wohnsitz):* die soziale, regionale M. der Arbeitnehmer; Bei der Arbeitssuche werden mehr M., geringere Lohn- und Gehaltsforderungen ... empfohlen (Saarbr. Zeitung 2.10. 79, 4); eine Gesellschaft mit hoher M. **3.** (Milit. selten) *mobiler Zustand, Kriegsbereitschaft:* eine Demonstration der hohen M. ... der sowjetischen Kriegsmarine (Bundestag 190, 1968, 10, 325).

4. Progress of the Project and Future Perspectives

Our dictionary and information system is a long term enterprise. The development of the information system follows roughly four phases, of which the first two have been completed.

In the conceptual phase, aims were defined and linguistic methods that are compatible with corpus-based approaches were developed. Var-

ious linguistic theories concerning the problem of ambiguity, vagueness and the arrangement of senses were examined. Furthermore, as a hypertext dictionary makes other forms of presentation possible, we evaluated a number of ways of presenting data, such as collocations, argument structure and paradigmatic structures, in order to move the project forward with illustrations of lexico-semantic structures of a quantity and quality that cannot be offered in a printed dictionary. During the process of developing the overall theoretical concept of *elexiko*, we were particularly concerned with finding a consensus about theoretical imperatives and lexicographic practice. Another central task was to establish efficient lexicographic procedures for the evaluation of mass data and, in particular, collocations.

During the second stage, the lexicographic data model (Document Type Definition) and further technological implementations, as well as the design of a data infrastructure, were developed and tested.

Phase three was launched in 2003 which comprises the actual compilation and writing of the dictionary. As a first milestone, the entire headword list, including first details such as variant spellings and syllabication has been publicly accessible since January 2004.[21] Accordingly, search options are still very limited. In addition, the headword list can be viewed alphabetically in reverse mode. As *elexiko* is being compiled following a modular approach, a batch containing about 250 lexemes and illustrating a particular semantic field (*Demonstrationswortschatz*) has been analysed and inserted into the database. By describing a semantic field, we were able to present a part of the lexicon with different types of lexical structures. This module will serve to demonstrate the potential of *elexiko* and can be accessed online since July 2004. At the same time, separate modules which describe multi-word entries for the *Demonstrationswortschatz* and neologisms of the 1990s are being incorporated into the *elexiko*-database and will be presented in 2005.

Whereas in printed editions of a dictionary each part of the microstructure has its fixed position and the reader is used to a specific order and linear arrangement of semantic and syntactic information, a hyper-

[21] See website of the project http://www.elexiko.de.

text dictionary has the advantage that the user decides what kind of information s/he is interested in. Within *elexiko* it is not necessary to read the entire entry, but to extract and present the required information from an entry selectively. The depth of information can be navigated according to individual interests, for example, by refining search options or suppressing portions of an entry. At the moment the search options are limited to searching for spelling variants and paradigmatic partners. The dictionary entries of *elexiko* are modelled as XML-instances, meaning that each piece of lexicographic information is marked up individually. The instances are stored in a content-oriented database with an underlying DTD that contains about 400 different tagging elements. In order to be able to extract specific information and in order to have the user's question answered (as outlined in the introductory part), a refined search and navigation system will be developed in the near future (phase four). So far, possible enquiries are very limited, but further query programming will allow us to go beyond the search options of electronic versions of other German dictionaries, such as *Duden-GWDS* (2000), *Duden-Universalwörterbuch* (2003) or *DWB-Online*. Once a complex search tool has been developed, *elexiko* will also serve as an information system where linguists can collect data for their own research. This targeted access of information also means that a diverse spectrum of users can explore the dictionary, no matter whether one's interest is professional, academic, or whether one is a foreign language learner or simply curious about the meaning or use of a particular German word.

The aforementioned principle of modularity can also be applied with regard to other projects that want to be connected to *elexiko*. A compatible hypertext structure is a prerequisite for becoming an integrated module. Initially, only internal IDS projects, such as the project 'neologisms of the 90s',[22] will be connected to *elexiko*. Such modules usually comprise a specific part of the German lexicon which might not be part of *elexiko* or which focuses on specific questions that are not covered by *elexiko*.

Further cooperation with external projects is desirable. Depending on the nature of collaborating modules, such cooperation could range

[22] See websites of the project: http://www.ids-mannheim.de/lexik/Neologie.

from being an integral part of *elexiko* to being merely a link to an external project. This networking vision will hopefully grow with the development of other online projects.

Acknowledgements

I wish to thank all the reviewers for their valuable comments.

5. References

Belica, Cyril. 1995-2002: *Statistische Kollokationsanalyse und Clustering. COSMAS-Korpusanalysemodul.* Mannheim: Institut für Deutsche Sprache. (http://corpora.ids-mannheim.de/cosmas).

Brisard, Frank et al. 2001: Processing Polysemous, Homonymous, and Vague Adjectives. In Hubert Cuyckens, Britta Zawada (ed), *Polysemy in cognitive linguistics*. Amsterdam: Benjamins, 261-284.

Cruse, Alan D. 1986: *Lexical Semantics*. Cambridge: CUP.

Cuyckens, Hubert/Zawada, Britta 2001: *Polysemy in Cognitive Linguistics*. Selected Papers from the Fifth International Cognitive Linguistics Conference Amsterdam 1997. Amsterdam and Philadelphia: Benjamins.

Haß-Zumkehr, Ulrike. 2001: *Deutsche Wörterbücher*. Berlin and New York: de Gruyter.

Haß-Zumkehr, Ulrike. 2004: Das Projekt Wissen über Wörter des Instituts für Deutsche Sprache‘ in Jürgen Scharnhorst (ed), *Sprachkultur und Lexikographie*. Frankfurt: Peter Lang.

Lutzeier, Peter Rolf 1981: *Wort und Feld.* Wortsemantische Fragestellungen mit besonderer Berücksichtigung des Wortfeldbegriffes. Tübingen: Niemeyer.

Moon, Rosamund. 1987: The Analysis of Meaning. In John Sinclair (ed), *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins, 86-103.

Polenz v., Peter. 1985: *Deutsche Satzsemantik*. Grundbegriffe des Zwischen-den-Zeilen-Lesens, (=Sammlung Göschen 2226). Berlin/New York: de Gruyter.

Pustejovsky, James 1995: *The Generative Lexicon*. Cambridge, Mass.: MIT Press.

Ravin, Yael/Leacock Claudia (eds) 2000: *Polysemy: Theoretical and Computational Approaches.* New York: Oxford University Press.

Reichmann, Oskar 1989: Einführung. In Robert R. Anderson, Ulrich Goebel and Oskar Reichmann (eds), *Frühneuhochdeutsches Wörterbuch. Bd.1, a – äpfelkern*. Berlin and New York: de Gruyter, 1-164.

Seiffert, Helmut 1969: *Einführung in die Wissenschaftstheorie*. Vol 1: Sprachanalyse - Deduktion - Induktion in Natur und Sozialwissenschaften. München: C. H. Beck.

Sinclair, John 1991: *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Storjohann, Petra 2003: The Lexicographic Use of Corpora and Computational Tools for Disambiguation. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds), *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL technical paper number 16. UCREL, Lancaster University.

Storrer, Angelika 1998: Hypermedia Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In Herbert Ernst Wiegand (ed), *Wörterbücher in der Diskussion III*. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. (Lexicographica. Series Maior 84.) Tübingen: Max Niemeyer, 105-131.

Storrer, Angelika 2001: Digitale Wörterbücher als Hypertexte: zur Nutzung des Hypertextkonzepts in der Lexikographie. In Ingrid Lemberg, Bernhard Schröder and Angelika Storrer (eds), *Chancen und Perspektiven computergestützer Lexikographie*. Hypertext, Internet, und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. (Lexicographica. Series Maior 107.) Tübingen: Max Niemeyer, 52-69.

Strauß, Gerhard 1989: Angabe traditioneller Wortarten oder Beschreibung nach funktionalen Wortklassen im allgemeinsprachigen Wörterbuch? In Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand and Ladislav Zgusta (eds), *Wörterbücher: ein internationales Handbuch zur Lexikographie*. (Handbücher zur Sprach- und Kommunikationswissenschaft Bd. 5.1). Berlin/New York: de Gruyter, 788-796.

Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): *Grammatik der deutschen Sprache*. 3 Bände, (IDS-Grammatik). Berlin / New York: de Gruyter.

Dictionaries

CIDE 1995 = *Cambridge International Dictionary of English*. Cambridge: CUP.

Duden-GWDS 1999 = *Das große Wörterbuch der deutschen Sprache*. 10 Bände. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Duden-GWDS 2000 = *Das große Wörterbuch der deutschen Sprache*. 10 Bände auf CD-Rom. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Duden-Universalwörterbuch 2003 = Duden - Deutsches Universalwörterbuch. CD-Rom. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

DWB 1854-1971/1984 = *Deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm*. 32 Bände - Online: http://www.dwb.uni-trier.de/index.html.

LDCE 1995 = *Longman Dictionary of Contemporary English*. London: Longman.

OALD 2000 = *Oxford Advanced Learner's Dictionary*. Oxford: OUP.