Published in: Proceedings of the Corpus Linguistics Conference 2005, Birmingham. - Birmingham: University of Birmingham, 2005. 20 Pp.

Corpus-driven vs. corpus-based approach to the study of relational patterns

Dr Petra Storjohann
Institut für Deutsche Sprache Mannheim (IDS)
storjohann@ids-mannheim.de

Abstract

Contextual lexical relations, such as sense relations, have traditionally played an essential role in disambiguating word senses in lexicography, as they offer insights into the meaning and use of a word. However, the description of paradigmatic relations in particular is often restricted to a few types such as synonymy and antonymy. The limited description of various types of relations and the method of presenting these relations in existing German dictionaries are often problematic.

Elexiko, the first German hypertext dictionary compiled exclusively on the basis of an electronic corpus, offers a new way of presenting sense relations, using a variety of approaches to extract the necessary data. In this paper, I will show how *elexiko* presents a differentiated system of paradigmatic relations including synonymy, various subtypes of incompatibility (such as antonymy, complementarity, converseness, reversiveness, etc.), and vertical structures (such as hyponymy and meronymy). Primary attention, however, will focus on the question of how data for a paradigmatic description is retrieved from the corpus. Whereas a corpus-driven approach is mainly used for various semantic information and a corpus-based method plays an important part in obtaining data for the grammatical description in *elexiko*, it will be argued that both the corpus-driven and the corpus-based approach can be complementary methods in gaining insights into sense relations. I will demonstrate which results can be obtained by each approach, and advantages and disadvantages of both procedures will be explored in more detail.

As sense relations are context-dependent, it will also be demonstrated how a sense-bound presentation can be realised in an electronic reference work including a system of cross-referencing that illustrates lexical structures and the interrelatedness of words within the lexicon. Finally, I will show how accompanying examples from the corpus and additional lexicographic information help the user to understand contextual restrictions, so that s/he is able to use dictionary information more effectively.

1 Preliminaries

The study of contextual relations, such as sense relations, is significant when investigating the structures of the lexicon of a language.

> Natural vocabularies are not random assemblages of points in semantic space: there are quite strong regularizing and structuring tendencies, and one type of these manifests itself through sense relations. (Cruse 2004: 143)

Sense relations offer insights into the meaning and use of a word, and they reveal the interrelatedness of the vocabulary. As Cruse (1986: 16) points out "the meaning of a word is fully reflected in its contextual relations". However, contextual relations not only possess a fascination for semanticists, but they also attract the interest of lexicographers. Contextual relations contribute to the semantic identity of a word, and they have therefore always played an important role in disambiguating word senses in lexicography (cf. Reichmann 1989: 111-114). The lexicographic treatment of paradigmatic structures, as one major type of sense relations, will be the focus of this paper.

Judging by the relatively large number of dictionaries that cover paradigmatic items (pairs, triplets, or more complex word sets), dictionary users have a strong interest in this type of information. Such dictionaries are consulted in specific situations of text production when a user searches for alternative expressions in order to specify, to generalize or simply to vary in style or register (cf. Wiegand 2004: 36). However, in many monolingual German dictionaries the description of paradigmatic relations is often problematic and limited to a few types, such as synonymy and antonymy, and their presentation is inadequate.

Paradigmatic patterns can illustrate specific semantic choices of a lexical item within a context, and their investigation can help to detect particularities of word meanings. A dictionary that aims at describing the meaning and the use of a lexical item should also include a semantic description of paradigmatic contextual partners, not only to illustrate the semantic identity of a lexical item but also to demonstrate the interdependency of words. As Hanks (1990: 35) argues:

> […] there is a tendency for human lexicographers to focus on the way words are used to describe the world rather then on the way words interrelate with one another.

With the availability of large computer corpora, paradigmatic contextual choices can be studied empirically, revealing selectional preferences and contextual constraints and conditions. Although corpora offer fundamental methodological advantages, corpus-assisted

approaches have, thus far, not played a central part in extracting and describing paradigmatic relations in German lexicography.

Elexiko is a relatively new lexicographic project based at the *Institut für Deutsche Sprache* in Mannheim (IDS) which aims to explain and document German and its present-day usage (cf. Haß-Zumkehr 2004, Storjohann 2005, and http://www.elexiko.de) including a detailed paradigmatic description of each lexical item. This electronic dictionary offers a differentiated presentation of sense relations and uses various corpus approaches to retrieve the necessary data. First, I will briefly outline the types of sense relations that are of interest to *elexiko*. Attention is then turned to the principal objective of this paper. I will explore how the required data for the paradigmatic description of a word is elicited from the corpus using a variety of methods. Finally, I will demonstrate how sense relations are presented lexicographically in *elexiko*.

2 The System of Paradigmatic Relations

The specificity of a lexeme's meaning in context can vary enormously. Following a contextual approach this meaning reveals itself through contextual relations. In order to account for a detailed description of the meaning and use of a word, lexical patterns, such as manifested paradigmatic sense relations, need to be examined. In *elexiko*, the illustration of paradigmatic patterns is part of the semantic description of a lexeme comprising the comprehensive demonstration of the horizontal and vertical relations which exist between the senses of lexical items (cf. lexical units in Cruse 1986: 84). These concern relations of inclusion and identity, as well as relations of exclusion and opposition. *Elexiko* has primarily adopted a classification following that offered by Cruse (1986) and by Lutzeier (1981), and this can be summarized as follows:

horizontal structures		**vertical structures**
incompatibility	antonymy	hyperonymy
	complementarity	hyponymy
	converseness	holonymy
	reversiveness	meronymy
synonymy		

Table 1

The major differences between this classification and paradigmatic categories in other existing German dictionaries (e.g. DUDEN 8, DUDEN WUG, WSA, WGDS, DORNSEIFF) concern the detailed distinction of terms of exclusion. The relations of contrast and

opposition, of which incompatibility is the most general sense relation, are divided into four categories. Whereas in other dictionaries the main relation of opposites is defined as antonymy, in *elexiko* (following Cruse 1986) this relation is a special case of incompatibility that is restricted to semantically gradable adjectives. Complementarity, converseness, and reversiveness are also specific sense relations of opposition and subtypes of incompatibility. Within vertical patterns, lexical relations are separated into hyponymy/hyperonymy and meronymy/holonymy. More precise definitions of individual relations, including specific types and subgroups, can be found in Cruse (1986 and 2004). Synonymy in particular is not further subclassified in *elexiko*, but is used to refer to all types of semantic identity, ranging from absolute sameness and propositional identity to more vague categories such as near-synonymy.

3 Corpus Retrieval of Sense Relations

As far as the lexicographic process of describing lexemes and their uses is concerned, the corpus is primarily being used exploratorily. Instances of natural language are studied in order to identify rules and patterns, and linguistic proto-typicalities are then interpreted and classified. Finding copious illustrative text samples is only a by-product of corpus-aided analysis. Besides an extensive and maximally representative corpus serving as an empirical basis, the lexicographic process of obtaining paradigmatic sense relations requires a good corpus query tool assisting the search of the corpus and processing data.

> Computers do not get bored; they notice only what they are told to notice; and they notice every occurrence of the word or usage pattern in the corpus that they have been told to notice, no matter how many there may be. Only a large corpus of natural language enables us to identify recurring patterns in the language and to observe collocational and lexical restrictions accurately. (Hanks 1990: 36)

However balanced the underlying corpus might be and however well the necessary software to search and analyse language data might work, another crucial prerequisite of good lexicographic work is the linguistic competency of data interpreting. Language data used for our lexicographic interpretation is retrieved exclusively from the *elexiko*-corpus, a monitor corpus currently comprising about 1,300 million words. For the extraction of paradigmatic partners, both the corpus-driven and the corpus-based approaches are applied (cf. Sinclair 1996 and Tognini-Bonelli 2001), as in practice, it was observed that an interplay of both methodologies can have substantial benefits for the retrieval of this type of sense relation.

3.1 The Corpus-Driven Approach

The corpus-driven approach (henceforth CDA) is a methodology whereby the corpus serves as an empirical basis from which lexicographers extract their data and detect linguistic phenomena without prior assumptions and expectations (cf. Tognini-Bonelli 2001). Any conclusions or claims are made exclusively on the basis of corpus observations. Unlike in English lexicography (cf. Sinclair 1987), this approach has, to date, not been employed in German lexicography. In *elexiko*, linguistic regularities within lexical relations are detected with the help of the computational analysis of collocations and the analysis and interpretation of concordances as found in the underlying *elexiko*-corpus. This corpus is searched with the corpus analysis tool COSMAS (Corpus Search, Management and Analysis System, http://www.ids-mannheim.de/cosmas2/) and the software package *Statistische Kollokationsanalyse und Clustering*,[1] an integral part of COSMAS, is used to process the retrieved data and to perform a collocation analysis.

The analysis of statistically significant co-selections of a lexical item enables the lexicographer direct access to lexical networks, among which sense relations of different kinds are often present. The computational analysis of collocations of a search item is, hence, the starting point for identifying paradigmatic relations. The following example – *flexibel* (27,424 instances) – will demonstrate the lexicographic procedure. In table 2 automatically retrieved paradigmatic collocates of *flexibel* are listed.

Total	**Anzahl**	**Autofokus**		**LLR**	**Kookkurrenzen**
		von	**bis**		
18576	422	-2	3	1314	**schnell**
10109	1	-2	3	1029	**mobil** dynamisch kreativ
10115	6	-2	3		mobil dynamisch
10118	3	-2	3		mobil kreativ
10120	2	-2	3		mobil Beschäftigte
10304	184	-2	3		mobil
14243	4	-3	3	607	**dynamisch** unternehmen
14247	4	-3	3		dynamisch mögen
14330	83	-3	3		dynamisch
14506	9	-2	3	587	**effizient** arbeiten
14507	1	-2	3		effizient billig
14511	4	-2	3		effizient Dienstleistung
14556	45	-2	3		effizient
14852	4	-4	6	579	**individuell** Wunsch

[1] The tool *Statistische Kollokationsanalyse und Clustering* was developed on the basis of statistical methods by Cyril Belica (1995-2002) at the IDS Mannheim and can be used free of charge online since 1995 (see also http://corpora.ids-mannheim.de/cosmas).

14912	60	-4	6		individuell
15350	1	-6	6	544	**starr** Modell ersetzen
15351	1	-6	6		starr Modell
15353	2	-6	6		starr ersetzen
15416	63	-6	6		starr
15711	4	-2	-2	497	**rasch** verändern
15770	59	-2	-2		rasch
19399	5	-4	3	447	**innovativ** unternehmen
19403	4	-4	3		innovativ handeln
19434	31	-4	3		innovativ
20146	2	-3	2	393	**kreativ** bleiben
20148	2	-3	2		kreativ motivieren
20181	33	-3	2		kreativ
21445	4	-2	-1	251	**modern** Dienstleistung
21512	67	-2	-1		modern
21865	3	-3	4	226	**intelligent** belasten
21883	18	-3	4		intelligent
21953	18	-4	-1	221	**klein** Einheit
22039	86	-4	-1		klein
22151	24	-4	3	211	**pragmatisch**
22683	1	-6	6	169	**sozial** einigen
22686	3	-6	6		sozial ab-
22705	19	-6	6		sozial
22875	10	-4	3	160	**transparent**
22896	4	-6	6	158	**anpassungsfähig**
23045	3	-3	-1	151	**offen** Gestaltung
23092	47	-3	-1		offen
23177	7	-5	6	147	**variabel**
23206	21	-3	3	143	**billig**
23282	8	-5	1	138	**beweglich**

Table 2

The data processing tool has been used to successfully exploit the corpus and thereby gain insights into the types of relations a search item enters into with other words in the same contextual environment. This is done without relying on intuition and personal linguistic competence. Investigations of collocations have shown that relations identified as significant, typical, and conventional often did not correspond with the expectations of the lexicographer. Therefore, CDA proves indispensable, since it provides information on significant and typical sense relations.

> There might be a large number of potentially meaningful patterns that escape the attention of the traditional linguist; these will not be recorded in traditional reference works and may not even be recognised until they are forced upon the corpus analyst by the sheer visual presence of the emerging patterns in a concordance page. (Tognini-Bonelli 2001: 86)

Although many paradigmatic terms can be ascertained directly through the collocation analysis, the lexicographer cannot omit his/her linguistic interpretation of the statistical findings. After the retrieval of collocates, potential paradigmatic terms are to be examined and their contexts are to be analysed in order to validate and classify sense relations. Thus, the lexicographer conducts an analysis of concordances of a corresponding collocate, in order to identify the kind of relation attested to in contextual use, in order to prepare information for lexicographic description, and to choose illustrative text samples.

However, terms that are related paradigmatically with the search item *flexibel* can also be found indirectly in more intricate syntagmatic patterns involving other statistically significant co-occurrences that are not paradigmatic collocates. For instance, the incompatibles *selbstständig* (independent)*, teamfähig* (*team-oriented*), and *vielseitig* (*versatile*), which also enter into sense relations with *flexibel* in the sense of 'anpassungsfähig' ('flexible, adaptable'), were not direct findings, as they were not classified as autonomous collocates. Rather they were gained indirectly through an investigation of other significant collocates, in this case verbal co-selections. The incompatible *vielseitig* was detected when interpreting the contexts and syntagmatic patterns of the verb *einsetzen*. Similarly, paradigmatically related words, such as *selbstständig* and *teamfähig*, were found in syntagmatic patterns of various verbal collocates (see table below).

verbal collocates	syntagmatic patterns
agieren	*flexibel und selbstständig agieren*
bleiben	*flexibel und vielseitig bleiben*
einsetzen	*kann flexibel und vielseitig eingesetzt werden*
gestalten	*flexibel und vielseitig gestalten*
sein	*flexibel und selbstständig sein* *flexibel und teamfähig sein*
werden	*flexibel und selbstständig werden*
zeigen	*sich flexibel und teamfähig zeigen*

Table 3

Generally, the corpus-driven approach offers two different results. On the one hand, direct results are ascertained from the computer analysis of collocation where a paradigmatic partner is a statistically significant collocate. On the other hand, indirect results are obtained where a sense relation is identified through the analysis of a collocation partner which itself is not a

paradigmatic, but a significant syntagmatic partner illustrating more complex syntagmatic structures and embedding further paradigmatic lexical relations.

Employing CDA leads to different results than using an approach based on introspection with regard to frequency and typicality of patterns. The holistic approach to the corpus is the major advantage of the CD method for tracing paradigmatic relations lexicographically.

> The unexpectedness of the findings derived from corpus evidence leads to the conclusion that intuition is not comprehensively reliable as a source of information about language. (Tognini-Bonelli 2001: 86)

An introspective approach is problematic, as the native speaker's personal knowledge or intuition is not directly accessible or observable. It cannot account for all possible contexts of a lexeme, nor trace all possible sense relation, and neither can it account for central and typical patterns of a paradigmatic term. As Sinclair (1997: 29) points out "the main organizing procedures for composing utterances are subliminal, and not available to conscious introspection." Introspection is, however, crucial for the interpretation of textual evidence, for the analysis of collocation results, and for the identification of lexical relations.

However advantageous the corpus-driven method can be when tracing the paradigmatics of most lexemes, it cannot provide a comprehensive description of the sense relational patterns in some cases. A variety of factors (see section 3.2) determine whether the exclusive employment of CDA is sufficient to elicit paradigmatically related words. It does not seem possible to generalise which words can be described in detail paradigmatically through CDA. In effect, all lexemes which are not very frequent can be examined in an exclusively corpus-driven way by analysing all concordances "manually". Certain catchwords (e.g. *Mobilität, Flexibilität, Urbanität* etc.) often co-select other catchwords in the immediate lexical neighbourhood and, hence, show a large number of incompatible sets. These are prone to being captured quickly by a collocation search. Furthermore, the syntactic behaviour of a lexeme can have an effect on collocation findings. This holds particularly true for verbs which often expose collocates that reflect typical thematic roles, but rarely present verbal paradigms in generated collocations.

Although one can derive valuable results from CDA, in a number of cases, it cannot provide a comprehensive description of paradigmatic structures. Here, the corpus-based approach is used complementarily.

3.2 The Corpus-Based Approach

The corpus-based approach (hereafter CBA) is a method that uses an underlying corpus as an inventory of language data. From this repository, appropriate material is extracted to support intuitive knowledge, to verify expectations, to allow linguistic phenomena to be quantified, and to find proof for existing theories or to retrieve illustrative samples. It is a method where the corpus is interrogated and data is used to confirm linguistic pre-set explanations and assumptions. It acts, therefore, as additional supporting material.

> In this case, however, corpus evidence is brought in as an extra bonus rather than as a determining factor with respect to the analysis, which is still carried out according to pre-existing categories; although it is used to refine such categories, it is never really in a position to challenge them as there is no claim made that they arise directly from the data. (Tognini-Bonelli 2001: 66)

Although through this approach pre-existing categories cannot be challenged and it cannot provide for unexpected findings, in *elexiko* it is sometimes used as an additional tool for the extraction of some paradigmatic items, particularly to extend or complete paradigmatic descriptions. It is a supplementary procedure which is applied in the following cases:

First, some paradigmatic words do not occur in the immediate lexical surrounding and are, hence, not captured by the computational analysis of collocations. They co-occur in a wider context, usually within a contextual proximity of one or two sentences. This mainly concerns synonyms, hyperonyms, hyponyms, and in fewer cases, also terms of contrast and opposition.

Secondly, particularly verbs which are characterized by syntactic valency often co-select nouns which reflect thematic roles. Therefore, computer-generated collocations often lack verbal paradigmatic terms. The following example – *akzeptieren* (*accept*) (67,439 instances) – will serve as an illustration.

Total	**Anzahl**	**Autofokus**		**LLR**	**Kookkurrenzen**	
		von	**bis**			
14364	65	1	1	8383	werden Kreditkarten	(credit cards)
18099	62	1	1	4177	wird Bevölkerung	(population)
21078	287	-1	-1	3210	allgemein	(general)
23153	142	-3	-1	2015	gesellschaftlich	(social)
25570	558	-5	3	1583	Entscheidung	(decision)
25980	6	-1	-1	1421	voll Bevölkerung	(population)
26313	99	-5	4	1354	Kreditkarten	(credit cards)
26611	284	-5	-1	1340	Bedingungen	(conditions)
27719	204	-4	-1	1212	Bevölkerung	(population)
27934	92	-5	4	1074	zähneknirschend	(reluctant)

28765	88	-1	-1	1010	stillschweigend	(tacit)
31162	235	-5	5	732	Vorschlag	(suggestion)
34663	155	-5	-1	480	Gesellschaft	(society)
34913	224	-5	3	474	Angebot	(offer)
34997	74	-5	4	470	Entschuldigung	(apology)
35154	52	-1	-1	460	widerwillig	(unwilling)
35428	4	-5	4	399	Kompromiß vorgeschlagenen	(compromise)
35521	93	-5	4		Kompromiß	(compromise)
35306	88	-1	-1	413	grundsätzlich	(fundamental)
38309	59	-5	4	352	Kompromiss	(compromise)
38911	38	-1	-1	318	einstimmig	(unanimous)
39070	2	-5	-1	310	Mehrheit Bürger	(majority)
39195	125	-5	-1		Mehrheit	(majority)
39441	2	-5	-1	305	Entscheidungen demokratische	(decision)
39508	67	-5	-1		Entscheidungen	(decision)
39510	2	-5	-1	302	Regierung demokratisch	(government)
39736	224	-5	-1		Regierung	(government)
47456	16	-2	-1	100	Kompromisse	(compromise)
45372	5	-3	-1	141	fraglos	(undoubtedly)

Table 4

Most collocates of *akzeptieren* indicate typical syntactico-collocational slots such as subjects (e.g. *Bevölkerung, Gesellschaft, Mehrheit,* or *Regierung*) and objects (e.g. *Angebot, Bedingung, Entscheidung, Entschuldigung, Kompromiss, Konditionen, Kreditkarte, Vorschlag*) or adjectival adjuncts (e.g. *allgemein, einstimmig, fraglos, gesellschaftlich, kampflos, stillschweigend, widerwillig ,* or *zähneknirschend*). Semantically related terms such as synonyms (e.g. *anerkennen, annehmen, billigen, dulden, hinnehmen, zulassen,* or *zustimmen*) or complementaries (*ablehnen* or *sich weigern*) cannot be identified by CDA.

Thirdly, some paradigmatic items which are not statistically significant are still of interest to dictionary users. For example, the complementaries of the lexeme *sozial – asozial* and *unsozial* – cannot be traced by CDA due to their relative lack of significance. Learners of German, however, would expect to find such semantic counterparts because their interest may lie precisely in the sense-specific use of a negation prefix.

Finally, for many ambiguous words one specific sense occurs frequently and tends to dominate other senses in the corpus: hence, automatically retrieved collocates can often be allocated to one specific sense only. Proportionally, some collocates are suppressed by statistics. Cases of ambiguous terms which have a sense restricted to a national variety, for example Swiss or Austrian German, are similarly problematic.

For all of the cases mentioned above, CBA offers an additional, complementary method of tracing paradigmatic pairs. The corpus-based approach implies a specific corpus inspection, where the lexicographer has a specific paradigmatic word in mind and searches the corpus for samples to either invalidate or verify and quantify the assumption. With the help of introspective expectation, through the collation of existing dictionaries and the use of specific search options, valuable evidence can be elicited from the corpus and incorporated into the paradigmatic description. To return to the example of *flexibel* provided for CDA, the following table illustrates the entire paradigmatic set of *flexibel* in the sense of 'anpassungsfähig' and demonstrates which supplementary information was gained through CBA.

Paradigmatic term	**Corpus-driven approach (CDA)**	**Corpus-based approach (CBA)**
synonyms	*anpassungsfähig, beweglich, variabel, wendig*	*elastisch, wendig*
incompatibles	*anpassungsfähig, beweglich, dynamisch, individuell, kreativ, mobil, offen, rasch, selbstständig, teamfähig, vielseitig*	
incompatibles	*effizient, schnell*	*kostengünstig*
antonyms		*stur*
complementaries	*starr*	*fest, kompromisslos, unbeweglich, unflexibel*

Table 5

Whereas statistical significance plays an important role in CDA, frequency is also regarded as necessary evidence when extracting paradigmatic terms through the CB-method in *elexiko*. A paradigmatic word is defined through CBA as one which occurs in several sources (at least three independent texts) and over a number of years.

As Tognini-Bonelli (2001) emphasises, that using CBA exclusively cannot offer a holistic and systematic approach to a corpus. Hence, the procedure described here is an additional, supplementary step for tracing paradigmatic items in *elexiko.* It is carried out after the employment of CDA, in cases where the paradigmatic description remains incomplete or extensible. In cases where the CD procedure offers a comprehensive and detailed description of sense relations, CBA is not applied as an additional method.

4 The Lexicographic Presentation of Sense Relations in *elexiko*

Elexiko currently contains 300,000 single-word entries with minimal information on spelling, syllabication and grammar and it has been publicly accessible via the Internet since 2004 (www.elexiko.de). Approximately 350 entries have been fully lexicographically described containing detailed semantic, pragmatic, grammatical, and diachronic information as well as information on morphology and word formation. *Elexiko* is characterized by continuous growth and changes; new entries are added daily.

In *elexiko*, information on sense-related words is found in a separate sense-bound rubric[2] labelled "Sinnverwandte Wörter".

Bedeutungs-erläuterung	Semantische Umgebung u. lexikalische Mitspieler	Typische Verwendungen	**Sinnverwandte Wörter**	Besonderheiten des Gebrauchs	Grammatik

Figure 1

Traditionally, synonyms or antonyms are listed for a word as a lemma. However, defining any semantic relation as a lemmatic relation is problematic because paradigmatic relations hold between lexical units together with their senses. 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. The only appropriate way to demonstrate paradigmatic structures is a context-dependent presentation. Paradigmatic word sets are given for the different senses (Lesart) and sub-senses (Spezifizierung) of a word in *elexiko* (see figure 2 *mobil* (*mobile*) 'nicht gebunden' ('not bound/fixed')).

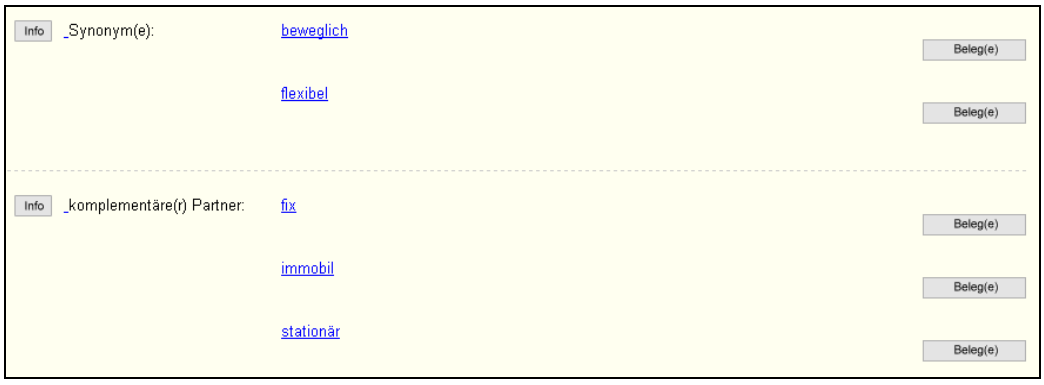

Figure 2

As most dictionary users are not familiar with terms such as *complementarity, incompatibility, reversiveness, converseness* etc., each semantic term is defined and illustrated with examples

[2] There are other dictionary rubrics referring to the lemmatic level, such as spelling, syllabication, word formation, diachrony, etc.

in a text that can be consulted via a separate box labelled Info. Whereas readers are used to traditional categories such as synonymy, dictionary users are not familiar with the term incompatibility, a sense relation holding between senses of lexemes which are often listed in onomasiological dictionaries (e.g. DORNSEIFF for German). The following example *Beruf* (*occupation*), together with its sense 'Arbeit' ('work'), demonstrates how this type of semantically related items plays a central part in identifying the discourse of the search item.

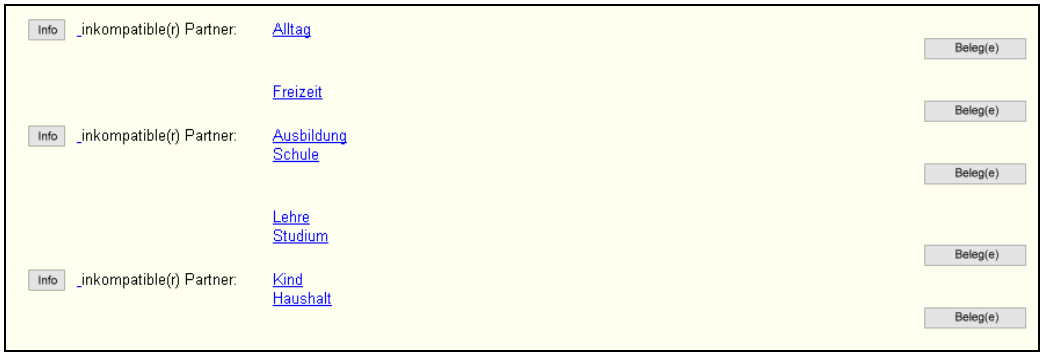

Figure 3

Cruse (2004) notes that often there is a set of incompatibles all of which have a common superordinate. Each set of incompatible terms refers to a specific notional area; they are confined to particular conceptual domains. In the example shown above (Figure 3) it can be seen that the terms of the set *Beruf – Alltag – Freizeit* (*occupation – everyday life – spare time*) denote a different conceptual context than the set which comprises the word set *Beruf – Ausbildung – Schule – Lehre – Studium* (*occupation – vocational training – school – university*), or contrastively the triplet *Beruf – Kind – Haushalt* (*occupation – child – home/household*).

4.1 Linking System

One of the major advantages of a hypertext dictionary is its capacity to include a large-scale hyperlinking system (mediostructure) for illustrating different types of language structures and to provide sufficient and quick cross-referencing. Incorporating hyperlinking enables the reader an instant follow-up and an improved perception of the interrelatedness of words within the lexicon. Within the section "Sinnverwandte Wörter" the linking system is an internal dictionary cross-reference. Each paradigmatic partner is presented as a hyperlink leading the user not only to the relevant entry, but also to the relevant sense or sub-sense which is attested for a sense relation with the corresponding lexical item. Given that the paradigmatic lexeme itself obtains a separate entry, it will be linked up systematically,

providing direct access to the interrelated item. Since the dictionary has not been fully compiled yet, linking is still very restricted. Only those words that have been fully described lexicographically are linked to the sense-describing level; others are provisionally linked to the lemmatic level that contains general information such as spelling and syllabication.

4.2 Corpus Samples

Exemplification is a key element in the presentation of sense relations in *elexiko*. Only a context constitutes a relation between concepts or between discrete sense units of lexical items, and therefore each relation is exemplified through an example illustrating the common contextual use. This serves several purposes. First, corpus samples are primary and actual evidence of the existence of the described lexical relation. Secondly, a corpus sample demonstrates the common conceptual ground of the paradigmatic pair. Thirdly, a given context can illustrate the semantic and syntactic embedding, that is, rules and constraints, as which govern usage, as the related items are shown in actual contextual use. And finally, since it is possible that a number of different sense relations can hold simultaneously between two lexical items, the incidence of multiple relations is attested by actual data evidence. This allows the user to compare different contextual relations directly and helps him/her to understand the contextual constraints which apply.

4.3 Additional Usage Information

Information on the usage of paradigmatic partners is rare in German monolingual dictionaries, or is restricted to details on register. In order to allow dictionary information to be used more effectively, it is possible to incorporate usage notes at any given point of information in *elexiko*. Within the paradigmatic description, this primarily means general notes (either labelled as Kommentar or Hinweis). Both serve different purposes and contain different types of information. Information contained in **Kommentar** comprises general additional lexicographic explanations and substantiations. On the other hand, **Hinweis** refers to specific restrictions of usage in context. Usage notes are written in informal German prose and in an explanatory style and are primarily provided in the following cases:

– multiple sense relations (e.g. *Heimat* ‘Zuhause’ with *Zuhause* as synonym and incompatible)

Figure 4

– regional restrictions of synonyms (e.g. *Fahrrad* ‘Fortbewegungsmittel’ and its Swiss synonym *Velo*)

Figure 5

– restrictions of synonyms according to the perspective of the speaker (e.g. *Korruption* ‘finanzielle Bestechung’ and its synonyms *Schmieren* and *Bestechlichkeit*)

Figure 6

– restrictions on semantic usages of synonyms (e.g. akzeptieren ‘anerkennen’ and its synonyms *dulden, hinnehmen*)

Figure 7

– syntactic particularities (e.g. *Anforderung* ‘Leistungsanspruch’ and its incomaptible *Erwartung* in their plural usage)

Figure 8

Lexicographic information is also provided in cases where a corpus observation has been made which runs counter to the lexicographer’s initial assumption and which has been retrieved via a corpus-based approach. Furthermore, if other dictionaries have been consulted and the given information cannot not be validated through the corpus, usage notes are sometimes provided. For instance, in the case of the entry *Kauf* ‘Bestechung’ a usage note emphasises that, contrary to other dictionaries, *Kauf* and *Bestechung* are not synonymous because they refer to different objects (inanimate vs. animate) and they denote processes which happen consecutively.

Kommentar:

Obwohl in verschiedenen Synonymwörterbüchern die Synonyme **Bestechung** und **Korruption** gebucht sind, lassen sich diese im elexiko-Korpus nicht in synonymischer Verwendung nachweisen. Ähnlich wie bei dem Verb **kaufen** in der Lesart 'bestechen', wird durch **Kauf** auf eine Handlung Bezug genommen, bei der eine Person etwas erwirbt, nachdem der Tatbestand der Korruption oder der Bestechung vorlag. Somit unterscheidet sich **Kauf** von **Bestechung** und **Korruption** dadurch, dass in gemeinsamen Kontexten ein anderes Objekt verwendet wird und erst nachdem z. B. die Bestechung eines Politikers stattfand, damit der Kauf von Wählerstimmen erfolgen konnte. In den Texten des elexiko-Korpus ist **Kauf** als Folge von **Bestechung** oder **Korruption** thematisiert. Darüber hinaus konnte anhand der Texte des elexiko-Korpus beobachtet werden, dass **Kauf** häufig, wenn auch nicht ausschließlich, ein unbelebtes Objekt verlangt, während **Bestechung** und **Korruption** belebte Objekte fordern.(Vgl. den folgenden Beleg.)

Bislang wurden 14 Personen aus dem olympischen Zirkel bekannt, die vom Bewerbungskomitee der Olympischen Winterspiele in Salt Lake City Zahlungen oder Leistungen im Wert von knapp 800 000 Dollar in Anspruch genommen haben. Die **Bestechung** der IOC-Mitglieder war laut einem Bericht der "New York Times" sogar von langer Hand geplant. Die der Zeitung vorliegenden Dokumente sollen beweisen, daß sich die Offiziellen der Olympia-Stadt 2002 schon 1991 darauf verständigt hatten, Gelder gezielt für den **Kauf** von Stimmen einzusetzen. (Berliner Zeitung, 01.02.1999, IOC weitet Untersuchungen aus, S. 35.)

Figure 9

5 Summary

Elexiko is the first German monolingual dictionary that extracts synonyms, incompatibles, hyperonyms etc. on the basis of a corpus. The comprehensive examination of paradigmatically related terms is guaranteed by a synthesis of two corpus-guided approaches. The interplay of two different corpus approaches in *elexiko* is labour intensive and requires considerable diligence on the part of the lexicographer, but it does frequently reveal general discrepancies between personal intuition and corpus evidence. As Hanks (1990: 40) correctly points out "natural languages are full of unpredictable facts […] which a corpus may help us to tease out". Paradigmatic structures which intuitively seemed common have proven to be unexpectedly uncommon: structures that were predicted as central or typical could not be verified or proved to be statistically insignificant.

Through the study of comprehensive corpus material, it is also possible to identify the limits of our theoretical framework. Some contexts cannot be allocated to one specific sense or sub-sense or a sense relation cannot be identified unequivocally. The investigation of corpus data has also revealed that there are sense relations which have not yet been fully described linguistically. Currently, research is being carried out to find solutions to some of theses problems. In the short term, we are looking for methods to determine further criteria for distinguishing sense relations and to find better and more user-friendly means of presenting sense relations, particularly with regard to visual illustrations. Although generally, sense relations have been studied in detail, corpus-guided investigations of relational patterns open up a number of new issues with respect to paradigmatic relations in actual text and discourse. Our work has shown that, combined with the new possibilities for presenting lexicographic information in an electronic medium, the subject of sense relations needs to be addressed, at least in part, from a different perspective.

6 References

Belica, C. (1995) *Statistische Kollokationsanalyse und Clustering*. Korpusanalysemodul. (Mannheim: Institut für Deutsche Sprache).

Cruse, A. (1986) *Lexical Semantics*. (Cambridge: Cambridge University Press).

Cruse, A. (2004) *Meaning in Language – And Introduction to Semantics and Pragmatics*. (Oxford: Oxford University Press).

Hanks, P. (1990) Evidence and Intuition in Lexicography, in J. Tomaszczyk, B. Lewandowska-Tomaszczyk (eds.) *Meaning and Lexicography* (Amsterdam/Philadelphia: Benjamins), 31-41.

Haß-Zumkehr, U. (2004) Das Projekt Wissen über Wörter des Instituts für Deutsche Sprache, in J. Scharnhorst (ed.) *Sprachkultur und Lexikographie. Von der Forschung zur Nutzung von Wörterbüchern* (Frankfurt: Peter Lang), 311-330.

Lutzeier, P. R. (1981) *Wort und Feld. Wortsemantische Fragestellungen mit besonderer Berücksichtigung des Wortfeldbegriffes* (Tübingen: Niemeyer).

Lyons, J. (1977) *Semantics* 2 vols. (Cambridge: Cambridge University Press).

Reichmann, O. (1989) Lexikographische Einleitung, in: R. R. Anderson, U. Goebel, and O. Reichmann (eds.) *Frühneuhochdeutsches Wörterbuch*, Vol. 1 Einführung, a-äpfelkern (Berlin/New York: Walter de Gruyter), 10-164.

Sinclair, J. (ed.) (1987) *Looking Up: An Account of the COBUILD Project in Lexical Computing* (London: HarperCollins).

Sinclair, J. (1996) The Search for Units of Meaning, in *TEXTUS*, Vol. IX, 75-106.

Sinclair, J. (1997) Corpus Evidence in Language Description, in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) *Teaching and Language Corpora* (London/New York: Longman), 27-39.

Storjohann, P. (2005) *elexiko* - A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics* 34.

Tognini-Bonelli, E. (2001): *Corpus Linguistics at Work* (Amsterdam/Philadelphia: Benjamins).

Wiegand, E. H. (2004) Lexikographisch-historische Einführung, in U. Quasthoff (ed.) *Dornseiff – Der deutsche Wortschatz nach Sachgruppen*. 8. Auflage. (Berlin/New York: Walter de Gruyter), 9-91.

Dictionaries

DORNSEIFF = *Der deutsche Wortschatz nach Sachgruppen*. 8., völlig neu bearbeitete und mit einem vollständigen alphabetischen Zugriffsregister versehene Auflage von Uwe Quasthoff.

Mit einer lexikographisch-historischen Einführung und einer ausgewählten Bibliographie zur Lexikographie und Onomasiologie von Herbert Ernst Wiegand (Berlin/New York: de Gruyter), 2004.

Duden 8 = Die sinn- und sachverwandten Wörter. Wörter für den treffenden Ausdruck. 2. Aufl. (Bibliographisches Institut Mannheim/Wien/Zürich: Dudenverlag), 1986.

Duden WuG = Wörter und Gegenwörter. Wörterbuch der sprachlichen Gegensätze. 2. durchgesehene Aufl. von Christiane und Erhard Agricola (Bibliographisches Institut Mannheim/Wien/Zürich: Dudenverlag), 1992.

WSA = Wörterbuch der Synonyme und Antonyme. Sinn- und sachverwandte Wörter und Begriffe sowie deren Gegenteil und Bedeutungsvarianten. Von Erich und Hildegard Bulitta. (Frankfurt: Fischer), 2003.

WGDS = Antonyme. Wörter und Gegenwörter der deutschen Sprache. Herausgegeben von Gudrun Petasch-Molling (Eltville: Bechtermünz), 1989.

Internet Ressources (accessed June 2005)

COSMAS = Corpus Search, Management and Analysis System: http://www.ids-mannheim.de/cosmas2/

elexiko = Das lexikalisch-lexikologisch korpusbasierte Informationssystem des IDS (www.elexiko.de)