Internet lexicography as a challenge: The Internet dictionary portal at the Institute for German Language

Stefan Engelberg, Annette Klosa and Carolin Müller-Spitzer

Stefan Engelberg is head of the department 'Lexik' at the Institute for German Language and professor for German linguistics at the University of Mannheim.
engelberg@ids-mannheim.de

Annette Klosa is manager of the project 'elexiko' at the Institute for German Language.
klosa@ids-mannheim.de

Carolin Müller-Spitzer is manager of the project 'OWID' at the Institute for German Language.
mueller-spitzer@ids-mannheim.de

1 Introduction

The Internet has become the major challenge for lexicography in the 21st century. Not only does it present lexicography with new possibilities of data integration and crosslinking, it has also changed the demands made on the competences that have to be gathered within lexicographic projects, reaching from the development of corpus analysis methods to text technology and web technology. In connection with these developments, the lexicographic work proper is becoming more complex: The ability to handle and analyse mass data and the need for Internet-adequate lexicographic concepts is changing the profession. In this article, we will take a look at the Internet dictionary portal developed at the Institut für deutsche Sprache ('Institute for German Language') in Mannheim and illustrate the lexicographic practice. In Section 2, we present the Online German Lexical Information System OWID. Section 3 illustrates the lexicographic practice and structure of *elexiko*, the largest dictionary published in OWID. In Section 4, the basic principles of data structuring and administration are discussed, and Section 5 gives an overview on the workflow and work organization within the department *Lexik,* which hosts the Internet-lexicographic projects of the Institute for German Language.

2 The lexicographic information system OWID

Dictionaries in OWID: The *Online-Wortschatz-Informationssystem Deutsch* (OWID; Online German Lexical Information System) is a lexicographic Internet portal for various electronic dictionary resources that are being compiled at the Institute for German Language (Institut für Deutsche Sprache, IDS; cf. OWID 2008ff.). The main emphasis of OWID is on academic lexicographic resources of contemporary German. Presently, the following dictionaries are included in OWID.

• "*elexiko*" (2003ff., cf. Haß 2005, Klosa et al. 2006) consists of an index of about 300,000 short entries with information on spelling, spelling variation, and word division. In addition, many entries contain citations automatically chosen from the *elexiko*-corpus. Furthermore, *elexiko* contains more than 1,000 fully elaborated entries of high-frequency headwords, focussing on extensive semantic-pragmatic descriptions of lexical items in actual language use. The dictionary is being extended continuously by further elaborated entries.

• The *Neologismenwörterbuch* (Dictionary of Neologisms) (2005ff., cf. Herberg et al. 2004) describes in detail about 800 new words and new meanings of established words added to the German vocabulary since the 1990s. This dictionary is also constantly upgraded.

• *Feste Wortverbindungen online* (Multiword Expressions Online) (2007ff.) publishes the research results of the project *Usuelle Wortverbindungen* (Fixed Multiword Combinations). Twenty-five detailed entries for fixed multiword combinations and 100 shorter entries dealing with additional collocations are currently available.

• The *Diskurswörterbuch 1945-55* (Discourse Dictionary 1945-55) (2007) is a reference work covering the lexemes that establish the discourse about "guilt" in the early post-war era 1945-1955. It subsumes the lexical-semantic results of the discourse-theoretical studies in Kämper (2005, 2007).

In the near future, the *Handbuch deutscher Kommunikationsverben*, a handbook of German communication verbs which consists of two volumes, a dictionary containing approximately 350 entries and a volume representing the lexical structures of German communication verbs by means of lexical fields (cf. Harras et al. 2004, Harras et al. 2007), the "VALBU - Valenzwörterbuch deutscher Verben" (Schumacher et al. 2004) (Valency Dictionary of German Verbs), and about 300 articles from a corpus-based project on proverbs will be published in OWID.

Access structure: The main function of OWID is to provide a common access structure in the form of search options across the individual dictionaries. This is the typical function of lexicographic portals (cf. Engelberg and Müller-Spitzer, forthcoming).

To give an example: if a user types "global*" in the search box of OWID, he/she gets the results displayed in Figure 1.

The entries from *elexiko* are presented in bold black, the entries from the neologism dictionary in bold black italics, those from the multiword dictionary (see below) in small capitals. (In the online presentation, information coming from

different dictionaries is rendered in different colours. Each dictionary is associated with a particular colour.) However, OWID is more than a meta search engine for the included dictionaries. The difference from other lexicographic portals is that the individual resources are explicitly interconnected to each other as can be seen in Figure 2.

Here, "blind" not only occurs as an *elexiko* entry, but is also part of the multiword combinations "blinder Aktionismus" ('blind action') and "blinder Alarm" ('false alarm') in the Collocations Online Dictionary. The two search results show that the lexicographic data is structured in a very granular way and strictly content based. For example, the search engine is able to output "globaler *Komparativ zu* **global**" or "Blind date *nichtnormgerechte Schreibvariante zu* ***Blind Date***"[1] (more about that topic in Section 3).

Dictionary versus portal: In OWID, there is a clear distinction between the level of the portal and the level of an individual dictionary. The display of the headword list illustrates this. Having used the search box on the OWID homepage, the entry requested is embedded in the entire OWID-headword list (that is, a joined headword list from all included dictionaries) (cf. Figure 3).

On the other hand, if a user chooses one individual dictionary by clicking on the dictionary button and looks up one entry, only the headword list from the individual dictionary is displayed (cf. Figure 4). With this differentiation, we meet two different user needs: firstly, searching for one word in no particular dictionary or, secondly, searching within one specific dictionary only.

Online bibliography OBELEX: Besides the main function as a lexicographic portal, OWID provides another service for researchers in the context of online lexicography, the "Online-Bibliography of Electronic Lexicography (OBELEX)." All publications recorded in OBELEX are cross-referenced by keyword and language (cf. Figure 5). Information on dictionaries is currently not included in OBELEX; the main focus is on metalexicography. However, we are working on a database with information on online dictionaries as a supplement to OBELEX.

3 Corpus-driven lexicography: the Internet dictionary *elexiko*

Corpus-driven lexicography: *elexiko* is a lexicological-lexicographic project compiling a reference work that explains and documents contemporary German.

It was specifically designed for publication

1 "nicht-normgerechte Schreibweise" 'non-standard spelling'.

· **global**
· Global Cities *Nominativ Plural von* ***Global City***
· global city *nichtnormgerechte Schreibvariante zu* ***Global City***
· Global city *nichtnormgerechte Schreibvariante zu* ***Global City***
· ***Global City***
· Global Citys *Nominativ Plural von* ***Global City***
· global player *nichtnormgerechte Schreibvariante zu* ***Global Player***
· global Player *nichtnormgerechte Schreibvariante zu* ***Global Player***
· Global player *nichtnormgerechte Schreibvariante zu* ***Global Player***
· ***Global Player***
· …
· globaler *Komparativ zu* **global**
· globalere *Komparativ zu* **global**

Figure 1 Results for the search term "global*" in OWID (with *Global City* and *Global Player* being English loanwords in German)

· **blind**
· blind *Basiselement zu* BLINDER AKTIONISMUS
· blind *Basiselement zu* BLINDER ALARM
· blind *Basiselement zu* BLIND DATE
· …
· blind date *nichtnormgerechte Schreibvariante zu* ***Blind Date***
· Blind date *nichtnormgerechte Schreibvariante zu* ***Blind Date***
· ***Blind Date***

Figure 2 Results for the search term "blind*" in OWID

Figure 3 OWID headword list and article from the dictionary of neologisms

Figure 4 Headword list and article from the dictionary of neologisms

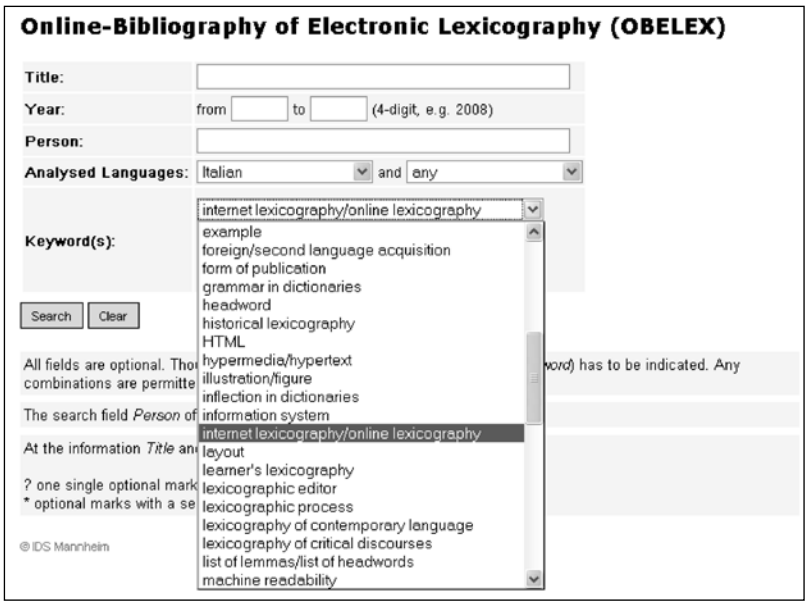

Figure 5 The Online Bibliography of Electronic Lexicography (OBELEX)

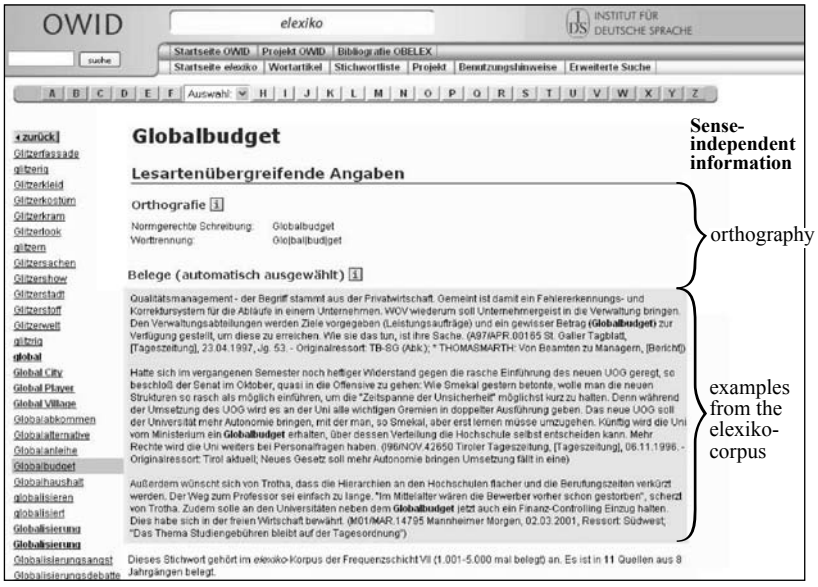

Figure 6 Entry **Globalbudget** with automatically compiled sense-independent information

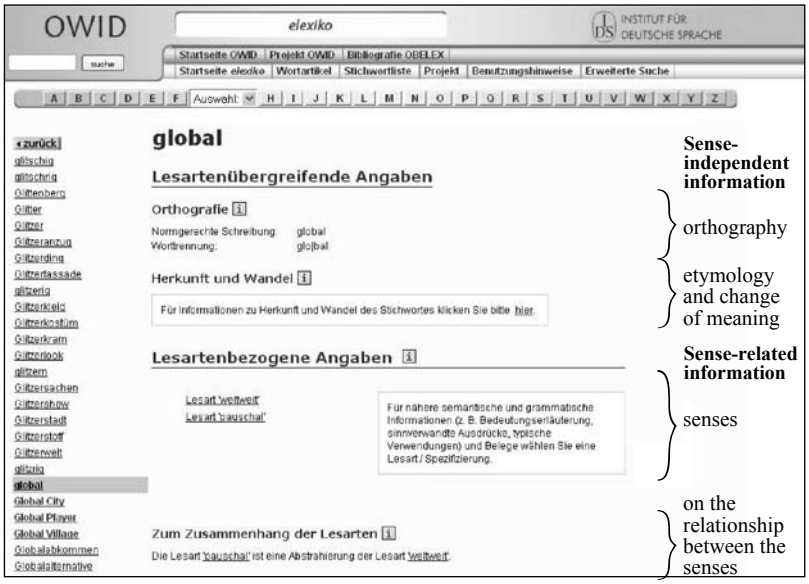

Figure 7 Entry **global** with information on the senses.

on the Internet and is published within the lexicographic information system OWID. If one refers to *elexiko* as an Internet "dictionary," one should keep in mind that *elexiko* is more than a dictionary in its traditional sense although, of course, it contains descriptions of the meaning and use of a lexeme just as any traditional dictionary. It is both, a hypertext dictionary and a lexical data information system (Storjohann 2005b, Klosa et al. 2006).

The primary and exclusive basis for lexicographic interpretation in *elexiko* is an extensive German corpus. The compilation of the dictionary does not proceed letter by letter. Instead, the headwords are grouped into word classes or word fields for the compilation process.

Modules may also be defined according to levels of frequency and distribution of lexemes in the *elexiko*-corpus. *Elexiko* is now working on a module called "Lexikon zum öffentlichen Sprachgebrauch" ("Dictionary on Public Discourse"). It contains approximately 2,800 entries selected mainly by their (high) frequency in the *elexiko*-corpus.[1]

The list of headwords[2] was taken exclusively from the *elexiko*-corpus and was published on the Internet before starting the work on lexicographic modules. For each headword, sense-independent information generated automatically or semi-automatically from the underlying corpus is given. This concerns 300,000 single-word entries comprising details on spelling, spelling variation, and word division. Many entries also contain automatically chosen citations from the *elexiko*-corpus (cf. Figure 6).

Lexicographically fully-described entries (as in the module "Dictionary on Public Discourse") entail sense-independent information on morphology and word formation ("Lesartenübergreifende Angaben") as well as a number of senses and their relationship. They also offer a large scope of sense-related information ("Lesartenbezogene Angaben") in detail: meaning definition, collocations, syntagmatic patterns, sense-related terms, pragmatics, and grammar.

For the process of writing and presenting *elexiko* on the Internet, the project uses numerous technologies and software tools, such as the corpus query and processing tool COSMAS II[3] developed at the IDS and its incorporated collocation

1 Examples given here are part of this module and may be viewed online.

2 For the list of headwords in alphabetical or reverse order, see http://www.owid.de/elexiko_/Stichwortliste.html/.

3 For COSMAS II, see http://www.ids-mannheim.de/cosmas2/.

software package "Statistical collocation analysis and clustering."[1] In addition, the corpus-linguistic research and development workbench CCDB[2] (also developed at IDS) is used. The corpus research software and the co-occurrence analysis are employed for numerous corpus-guided investigations within the practical working procedure.

Sense disambiguation: To disambiguate a highly frequent polysemous word in *elexiko,* we have "developed a disambiguation technique which is based on empirical and theoretical grounds. The lexicographic prerequisites of this disambiguation procedure are an elaborate theory, corpora, a data-processing software, and the linguistic competency of data interpreting." (Storjohann 2003: 755). Employing COSMAS II and the co-occurrence analysis, the lexicographer disambiguates a polysemous word at three levels: the collocation level, the KWIC-level, and the text level and, thus, achieves a detailed semantic description (Storjohann 2003: 755). In a second step, the senses found by a strictly corpus-driven method are contrasted to the senses given in other dictionaries. If the lexicographer finds that in *elexiko* the range of senses seems yet to be incomplete, the corpus will be searched again. In this way, corpus-driven and corpus-based information on all senses of a lexical item is combined.

All senses are listed in *elexiko* in a frequency-based hierarchy: The most frequent sense in the *elexiko*-corpus comes first. Senses in *elexiko* belonging to each other such as the literal and the metaphoric sense are, thus, sometimes torn apart. At the same time, there may be word senses that are not related to the others at all. Accordingly, in the *elexiko*-entries, users find an explanation on how the senses are arranged and how they are connected. On the first screen opening after typing in the search word, there is a list of all senses, which are not numbered but labelled with words (cf. Figure 7); here 'weltweit' ('worldwide') and 'pauschal' ('indiscriminate'). Under the rubric "Zum Zusammenhang der Lesarten" ("On the relationship between the senses"), their relationship is explained: The word sense 'pauschal' is an abstraction of the word sense 'weltweit.'

Sense-related information: When clicking on one of the senses from the list shown in Figure 7, the user opens the complete range of sense-related information ("Lesartenbezogene Angaben") in *elexiko*: meaning definition ("Bedeutungserläuterung"), collocations ("Semantische Umgebung und lexikalische Mitspieler"), syntagmatic patterns ("Typische Verwendungen"), lexical relations ("Sinnverwandte Wörter"), pragmatics ("Besonderheiten des Gebrauchs"), and grammar ("Grammatik"). All information is extracted from the corpus, either in a corpus-driven or a corpus-based approach (cf. Haß 2005, Klosa 2007).

The information on collocations given in *elexiko* is an example of corpus-driven information: Typical, highly frequent words co-occurring with the headword are extracted from the corpus (using the software package "statistical collocation analysis and clustering") and are arranged according to their semantic function by the lexicographer. The headword **global** in the sense 'weltweit' ('worldwide'), for example, is used to specify actions or to characterize issues. The entry in *elexiko* gives statistically significant partner words as shown in Figure 8 (*Was macht man global? agieren*, *denken*, … 'What can be done globally? to act, to think, …'; *Was ist global? Denken, Ebene,* …, 'What is global? thinking, level, …').

Besides characterizing syntagmatic semantic relations, *elexiko* also offers a new way of presenting paradigmatic sense relations in "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)" (Storjohann 2005a: 1, cf. also Storjohann 2005b). For example, the entry **global** 'weltweit' lists as synonyms *international* ('international'), *universell* ('universal'), *weltumspannend* ('global'), and *weltweit* ('worldwide'). The complementary partner words *lokal* ('local'), *provinziell* ('provincial'), *regional* ('regional'), and *national* ('national') are arranged into semantic groups as shown in Figure 9. Each of the partner words is illustrated in *elexiko* by a citation, and each one includes a hyperlink to the corresponding article. Other ways of presenting paradigmatic partners (e.g., in a semantic net) may be developed in the future.

In order to gather lexical information of this kind, the statistically significant co-selections of a word are analyzed with the corpus research software COSMAS II. Among the results of this analysis, the lexicographers will often find words related to the headword in a semantic way. This is why analyzing collocations automatically is the first step the *elexiko*-lexicographers take

Figure 8 Collocations for the entry **global** ('weltweit')

1 For the software package "Statistical collocation analysis and clustering," see http://www.ids-mannheim.de/kl/projekte/methoden/ka.html/.

2 For CCDB, see http://corpora.ids-mannheim.de/ccdb/ and Keibel and Belica (2007).

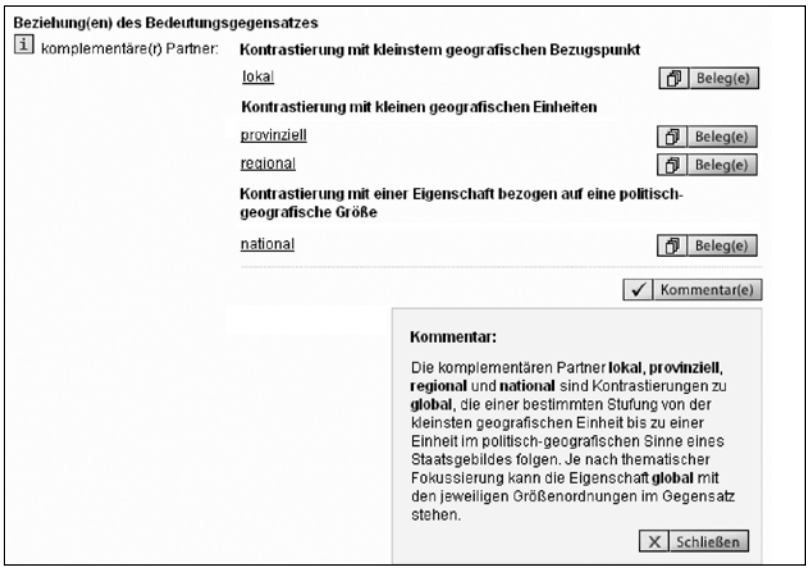

Figure 9 Groups of complementary partner words for **global** 'weltweit' ('worldwide')

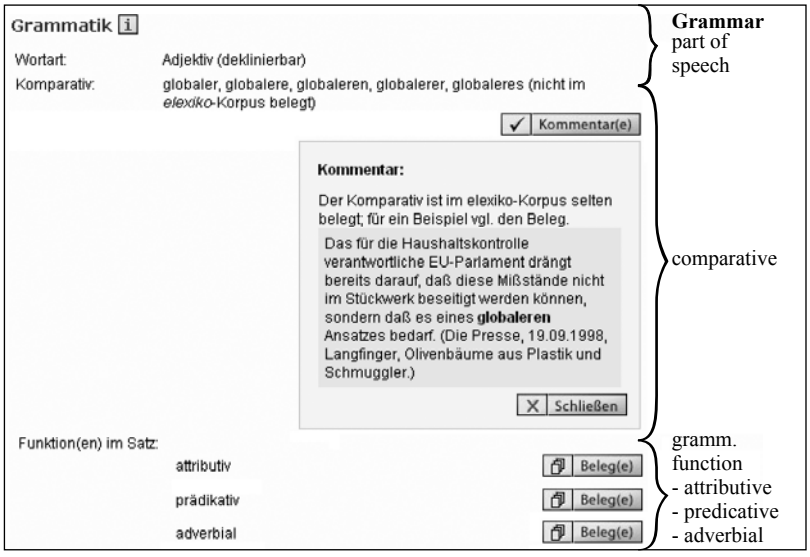

Figure 10 Grammatical information on **global** 'pauschal'

```
<adj-syntax>
 <adj-geltbereich>
 <adj-attributivA stellung="praenominal">
          <angabe-zusatz><belege><belegtextA>Die im Zuge der
          <belegwortA>globalen</belegwortA> Erwärmung zu
          erwartenden Klimakatastrophen werden in Europa die
          Gebiete im Süden und Osten härter treffen als jene im
          Norden. </belegtextA>
      <belegnachweisA>…</belegnachweisA></belege></angabe-zusatz>
 </adj-attributivA>

 <praedikativA>
          <angabe-zusatz><belege><belegtextA>Der Aggressor in
          diesem Krieg ist kein Staat. Es sind sich religiös
          definierende Terroristen, deren Basen, Operationsfelder
          und Zielgebiete <belegwortA>global</belegwortA> sind.
          </belegtextA>
      <belegnachweisA>…</belegnachweisA></belege></angabe-zusatz>
 </praedikativA>

 <adverbialA>
          <angabe-zusatz><belege>Weder die Öffnung gegenüber Europa
          oder der UNO noch Veränderungen im Innern sind selbst für
          Wirtschaftsführer, die <belegwortA>global</belegwortA>
          denken und handeln und gigantische Fusionen durchziehen,
          ein Thema.
      </belege></angabe-zusatz>
 </adverbialA>

</adj-geltbereich>
 </adj-syntax>>
```

Figure 11 Part of the XML file of the article global in *elexiko*, representing the syntactic functions of the adjective

on their way to identifying paradigmatic relations - without relying on intuition or personal linguistic competence. In a second step, the lexicographers evaluate the results gathered automatically: They classify the sense relations found and search for citations from the corpus texts that exemplify the sense relation concerned. Since in some cases the corpus-tools cannot provide a comprehensive description of the sense relational patterns, the corpus is checked based on a comparison with other dictionaries in a third, supplementary procedure, particularly so as to extend or complete paradigmatic descriptions.

Grammar in *elexiko*: Grammatical information in *elexiko* is also based on the corpus, and it is given for each sense. For adjectives, besides naming the part of speech, users find information on comparison and on attributive, adverbial, or predicative use (see Figure 10). All uses are only given if they are found in the corpus. Corpus-based grammatical information in *elexiko* is, thus, more detailed and more reliable than information given in many other dictionaries. On the other hand, a strictly corpus-based approach may cause problems: "When you analyze corpus data, you constantly find that inflectional forms given in the dictionary are in fact not attested in actual text productions, or the opposite situation of inflectional forms occurring that are not part of the official paradigm. The problem is primarily one of interpretation: is the absence of an inflectional form an indication that it does not exist, or is it an indication that the corpus is simply not large enough?" (Trap-Jensen 2002, cf. also Klosa and Müller-Spitzer 2007).

As can be seen from the information on comparison in Figure 10, each comparative form for the headword **global** is checked in the *elexiko*-corpus, and, if not found there, this is recorded ("nicht im *elexiko*-Korpus belegt", i.e., "no evidence in the *elexiko*-corpus"). In a commentary accompanying the comparative forms, the lexicographer may note the rare use in the corpus and give an example from the corpus texts.

4 Data structures and administration

Dynamic customizable microstructures: With respect to their contents, the individual participating projects and their compiled lexicographic resources in OWID are independent of each other. However, it has been obvious from the very beginning that the value of OWID would be increased if more common access structures for the different contents could be developed and if the lexicographic data would be interlinked more adequately. Above all, we

wanted to respect requirements of modern lexicography and dictionary research. For example, the dictionary user interface should be adaptable to specific dictionary consulting situations by creating dynamic customizable microstructures. "It is one thing to be able to store ever more data, but another thing entirely to present just the data users want in response to a particular look-up." (De Schryver 2003: 178, cf. also Engelberg and Lemnitzer 2001, Storrer 2001, and on the modelling concept Müller-Spitzer 2007) So on the one hand, in order to create a basis for a common access structure to the content, consistent principles for modelling and structuring the contents were applied to all integrated products. On the other hand, OWID will also be kept open for the possible integration of externally developed lexicographic resources, namely, reference works that are written outside IDS and other lexicological resources.

Data modelling: The approach chosen here not only guarantees that different lexicographic products can be integrated under the management of OWID on the macro structure level, that is, the level of headwords, but that dictionaries can also be accessed on a more granular level. Therefore, the attempt was to harmonize modelling on the level of content structure, that is, the level of the individual lexicographic information.

Technical architecture of OWID: OWID uses a single modelling process for all projects: All lexicographic data are stored as XML files. For each individual resource, a special tailor-made XML-DTD/ XML-schema was developed. In these DTDs, the microstructures of the individual dictionaries are defined. Focusing on the interconnectedness of the individual projects, a modular system was established where identical phenomena were modelled identically and only once. The dictionary entries are then written with an XML editor and stored in an Oracle database system. For the Internet presentation, the XML data are transformed by an XSLT stylesheet to HTML. To provide a uniform structure for lexicographic information of the same type contained in different dictionaries, a DTD library was created for OWID, where specific DTDs contain all entities, elements, or attributes that are shared by all entry structures. Due to this segmentation, the modelling level already shows what information is accessible across the different dictionaries. This procedure requires each individual information unit to be granularly tagged in all entry structures, but it also allows for automatic access to each content unit. Figure 11 provides an initial impression of the overall tagging granularity. It shows a part of the XML file of the entry **global** from *elexiko* (corresponding to the bottom part of Figure 10), illustrating the tagging of information concerning the grammatical functions of adjectives.

Editing system EDAS: Our internal editing system EDAS (Electronic Dictionary Administration System) allows lexicographers to use the granular XML structures for very special search questions with XPath, an expression language used to access or refer to parts of an XML document. On the basis of XML structures as in Figure 11, it is, for example, possible to search for all adjectives that are used attributively in prenominal position (cf. Figure 12).

As a result, adjectives from *elexiko* and the neologism dictionary are displayed. Searches on the content of paraphrases are possible as well. Figure 13 shows the results of a query that requests all entries that have the word "Computer*" in their paraphrase. It can be seen that all entries displayed in this search result belong to computer-related vocabulary.

Search options: These search options can only be used internally by the lexicographers. For the dictionary user, the search potential hidden in these data structures has only been partly revealed so far. However, the search option "Erweiterte Suche" (Advanced Search) within each dictionary already facilitates detailed searches for specific information. It is, for instance, possible

- to search in *elexiko* for all nouns with an old spelling variant that are compounds;

Figure 12 XPath search in EDAS (extract), displaying the list of adjectives corresponding to the restriction expressed as a search term

Figure 13 XPath search in EDAS (extract), displaying the list of headwords that include "Computer*" in their paraphrase

• to search in the dictionary of neologisms for all new lexemes (Neologismentyp="Neulexem") that entered the German language in the early 1990s (Aufkommen="Anfang der 90er Jahre"). Search results are words like *wegzappen* 'to zap to another channel', *Neufünfland* 'New Laender', or *abspacen* (partly calque of 'to space out');
• to search in the dictionary of neologisms for all verbs (Wortart="Verb") that gained a new sense (Neologismentyp="Neubedeutung") in the 1990s. (Results are verbs like *blicken* (new sense 'to understand something') or *surfen* (new sense 'browsing through the Internet').

These examples show only some of many possibilities. Similar searches can be defined for all approximately 450 possible elements and their additional attributes available within the OWID modular entry structures.

OWID and the user: It is our goal to provide a maximum of flexibility for the user interface of OWID. Therefore, the data are modelled solely with respect to the content; aspects of presentation and dictionary use are kept apart from considerations about data modelling. Thus, the same data can be displayed differently for numerous user types and look-up situations with no need to transform it. However, until now for electronic lexicography, it has not been systematically investigated in large empirical studies which functionalities are useful for particular user groups and situations. Therefore, we are running an academic project focused on user research and data relations within and between lexicographic resources. For OWID, today's challenge is - besides the continuous extension and enhancement of the individual dictionaries and integration of new dictionaries - to provide the user with an increasing range of more flexible display possibilities of this machine-readable puzzle, which can lead to new forms of using lexicographic information. In this manner, OWID is a living system that is modified and expanded continuously.

5 Competence and workflow

Technological and lexicographic competences: In order to compile Internet dictionaries like *elexiko* and create dictionary portals like OWID, a wide range of competences are required. The tasks that make up our Internet-lexicographic enterprise fall into four major domains: (I) corpus linguistics, (II) lexicography proper, (III) text-technology and computer lexicography, and (IV) metalexicographic research.

(I) ***Corpus linguistics***: All our lexicographic projects are based on corpora. This requires the constant acquisition and maintenance of large corpora of present-day German. Corpus research software ("COSMAS II") is developed within the institute as well as methods of corpus analysis, in particular methods based on co-occurrence analyses. Most of these developments are made available to the public via Internet. Other methods have been developed mainly for internal use, for example, the semi-automatic compilation of frequency-based lemma lists, methods for extracting and representing frequency changes in word use, and the automatic insertion of lexicographic examples from corpora. These activities require competences in the domain of statistics, text technology, data processing, etc.

Most of the work within the area of corpus linguistics is done in projects outside the department *Lexik* (reference corpora, co-occurrence analysis, corpus research software), in particular in the project group "Corpus Linguistics"[1] and the data processing division. The development of methods for the automatic extraction of lexicographic information from corpora and the acquisition of historic corpora are activities located within the department. These will be intensified in the future to meet our particular lexicographic needs.

(II) ***Lexicography proper***: Despite the indispensable corpus linguistic, text technological, and computer lexicographic

1 See http://www.ids-mannheim.de/kl/projekte/methoden/.

activities, most of the work falls into the domain of lexicography proper. As described in Section 2, this involves the application of corpus linguistic methods, the analysis of the results, and the compilation of lexicographic articles. It also involves a large amount of conceptual work and accompanying activities such as the maintenance of an elaborate editorial manual, as in the project *elexiko*. The work requires experienced lexicographers and linguists and the ability for sophisticated use of corpus analysis methods. Since there is still a lack of institutionalized training and study programs for lexicographers[1], we invite students for short periods of practical training, employ student researchers, and try to integrate lexicographers in the making into our lexicographic projects.

(III) ***Text technology and computer lexicography***: As described in Section 4, the lexicographic data are created in XML form, stored in an Oracle database, and then transformed by XSLT stylesheets into the HTML surface representations visible to the user. In addition, web-based methods of research on dictionary use are developed. Thus, the competences required range from text technology over database management to web technology.

These core competences are covered within the department. Additional support comes from the grammar department that is experienced in Oracle database application. It is responsible for the development of the editorial system EDAS and the generic bibliography system on which OBELEX is based.

(IV) ***Metalexicographic research***: Finally, the department is engaged in numerous types of metalexicographic research: Its main focus is currently research on dictionary use, on the linking of lexicographic data, and on the investigation and lexicographic representation of lexical semantic relations, in particular with respect to *elexiko* and OWID. Apart from the core lexicographic, linguistic, and corpuslinguistic knowledge, the research on dictionary use in particular requires competences about research methods within the domain of empirical social sciences and Internet-based user research. Besides these activities, the OBELEX bibliography is developed in OWID.

Personnel and workflow: Figure 14 gives an overview on the different activities that are involved in creating our Internet portal and dictionaries. It also illustrates which core activities are carried out within the department and which support is necessary from other departments within the Institute. Read from bottom to top, the chart also gives a rough description of the workflow. Many employees are involved in various tasks and projects, but most of them allocate only part of their time to activities closely related to our Internet lexicographic projects. Furthermore, most of them are involved in more than one of the above-mentioned four task domains. Altogether, the equivalent of 10-12 full-time positions is allocated to Internet-lexicographic activities within the department. Since a number of researchers have part-time positions, up to 18 lexicographers, linguists, text technologists, etc. are involved in our Internet-lexicographic projects. This does not include the support given by approximately 8 student researcher positions. The distribution of

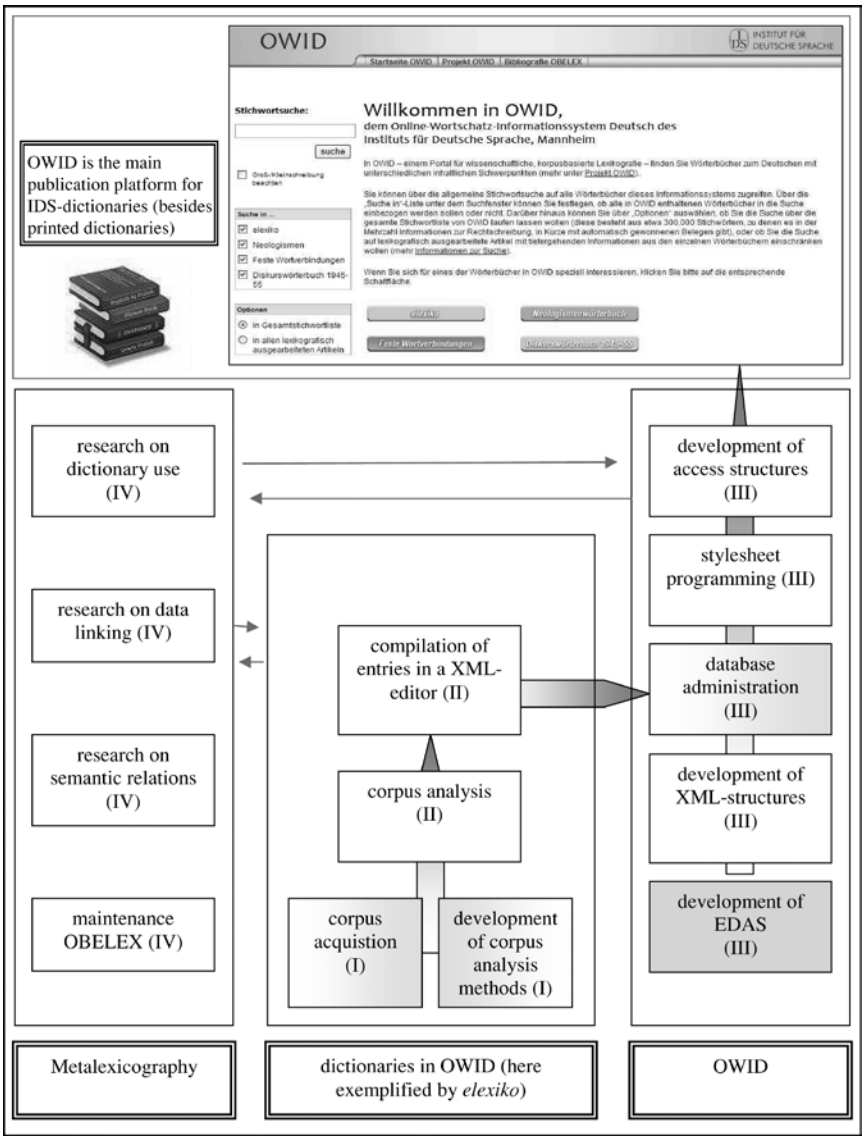

Figure 14 Task domains with regard to Internet lexicography (I = corpus linguistics, II = lexicography proper, III = text technology and computer lexicography, IV metalexicography) and an indication of the work flow. The shaded areas characterize the work done outside of the department Lexik

1 However, a "European Master in Lexicography" is currently being developed.

Coordination and conception of Internet lexicography	10 %
Task domain I: Corpus linguistics (in particular, automatic extraction of lexicographic information)	5 %
Task domain II: Lexicographic work proper (half of it in *elexiko*, the other half in the other dictionary projects mentioned in Section 2); including other activities like teaching classes, workshops on corpus-based lexicography, and Internet lexicography, etc.	50 %
Task domain III: Text technology and computer lexicography (DTDs, stylesheets, Oracle)	20 %
Task domain IV: Metalexicographic research (as described above)	15 %

Figure 15 The approximate distribution of positions within the department *Lexik*

positions within the department Lexik among our main Internet-lexicographic task domains appears in Figure 15.

Most of these positions are financed by the regular budget of the Institute, some of them - in particular, those allocated to metalexicographic research, the automatic extraction of lexicographic data from corpora, and the compilation of the dictionary of proverbs - are financed also with funds acquired from third-party donors.

6 Outlook

Internet lexicography will remain a main task for the department *Lexik*. The challenges posed by a thorough corpus orientation will continue to shape our lexicographic work in the future. This also holds for the conceptual demands that accompany the step from print to Internet lexicography, as is evident in projects like *elexiko*.

OWID will be the main platform for our lexicographic work. We are currently preparing the integration of several other dictionaries into OWID. However, for those dictionaries that were not modelled as fine-grained XML structures from the beginning, the conversion into proper XML structures is still tedious work. With the experience we are gaining, these tasks are becoming easier to carry out. Also, new dictionary projects are more likely to begin with adequately modelled XML structures from the start. New OWID dictionaries also pose new questions with respect to how to link the data and present them in a user-friendly and linguistically interesting format. Thus, besides broadening the lexicographic basis by integrating new dictionaries, the main and most interesting task will be to integrate the data in a way that goes beyond the creation of a common external access structure.

Bibliography

De Schryver, G.-M. 2003. 'Lexicographer's Dreams in the Electronic-Dictionary Age.' *International Journal of Lexicography* 16.2: 143-199.

Diskurswörterbuch 1945-55. 2007. *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim, Institut für Deutsche Sprache. www.owid.de/Diskurs1945-55/index.html/.

elexiko. 2003ff. *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim, Institut für Deutsche Sprache. www.owid.de/elexiko_/index.html/.

Engelberg, S., and Lemnitzer, L. 2001. *Lexikographie und Wörterbuchbenutzung* (= Stauffenburg Einführungen, Band 14). Tübingen: Narr.

Engelberg, S., and Müller-Spitzer, C. Forthcoming. 'Dictionary portal' in R.H. Gouws, U. Heid, W. Schweickhard, and H.E. Wiegand (eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin / New York: de Gruyter.

eValBu. 2009ff. 'elektronisches Valenzwörterbuch deutscher Verben' in *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache.

Feste Wortverbindungen. 2007ff. *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. www.owid.de/Wortverbindungen/index.html/.

Harras, G., Winkler, E., Erb, S., and Proost, K. 2004. *Handbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch* (= Schriften des Instituts für Deutsche Sprache 10.1). Berlin / New York: de Gruyter.

Harras, G., Proost, K., and Winkler, E. 2007. *Handbuch deutscher Kommunikationsverben. Teil 2: Lexikalische Strukturen* (= Schriften des Instituts für Deutsche Sprache 10.2). Berlin / New York: de Gruyter.

Haß, U. (ed.) 2005. *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. (Schriften des Instituts für Deutsche Sprache), Berlin / New York: de Gruyter.

Herberg, D., Kinne, M., and Steffens, D. 2004. *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen. Unter Mitarbeit von E. Tellenbach und D. al-Wadi* (Schriften des Instituts für Deutsche Sprache 11). Berlin / New York: de Gruyter.

Kämper, H. 2005. *Der Schulddiskurs in der frühen Nachkriegszeit. Ein Beitrag zur Geschichte des sprachlichen Umbruchs nach 1945* (Studia Linguistica Germanica 78). Berlin / New York: de Gruyter.

Kämper, H. 2007. *Opfer - Täter - Nichttäter. Ein Wörterbuch zum Schulddiskurs*

1945-1955. Berlin / New York: de Gruyter.

Keibel, H., and Belica, C. 2007. 'CCDB: A Corpus-Linguistic Research and Development Workbench' in *Proceedings of Corpus Linguistics 2007, Birmingham.* (http://www.corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf).

Klosa, A. 2007. 'Korpusgestützte Lexikographie: besser, schneller, umfangreicher?' in Kallmeyer, W., and Zifonun, G. (eds.), *Sprachkorpora - Datenmengen und Erkenntnisfortschritt* (Jahrbuch des Instituts für Deutsche Sprache 2006). Berlin / New York: de Gruyter, 105-122.

Klosa, A., and Müller-Spitzer, C. 2007. 'Grammatische Angaben in *elexiko* und ihre Modellierung' in Gottlieb, H. and Mogensen, E. (eds.), *Dictionary Visions, Research and Practice. Selected Papers from the 12th International Symposium on Lexicography, Copenhagen 2004*. Amsterdam: Benjamins, 13-37.

Klosa, A., Schnörch, U., and Storjohann, P. 2006. 'ELEXIKO – A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache, Mannheim' in Marello, C. et al. (eds.), *Proceedings of the 12th EURALEX International Congress (Atti del XII Congresso Internazionale di Lessicografia), EURALEX 2006, Turin, Italy, September 6th – 9th, 2006. Vol. 1.* Turin: Edizioni dell'Orso Alessandria, 425-430.

Kommunikationsverben online. 2009ff. *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache.

Müller-Spitzer, C. 2007. *Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis* (Studien zur deutschen Sprache 42). Tübingen: Narr.

Neologismenwörterbuch. 2005ff. *OWID – Online Wortschatz-Informationssystem Deutsch*. Mannheim: Institut für Deutsche Sprache. www.owid.de/Neologismen/index.html/.

OWID – Online-Wortschatz-Informationssystem Deutsch. 2008ff. Mannheim: Institut für Deutsche Sprache. www.owid.de/.

Schumacher, H., Kubczak, J., and Schmidt, R. 2004. *VALBU – Valenzwörterbuch deutscher Verben*. Tübingen: Narr.

Storjohann, P. 2003. 'The lexicographic use of corpora and computational tools for disambiguation' in Archer, D., Rayson, P., Wilson, A., and McEnery, T. (eds.), *Proceedings of Corpus Linguistics 2003 Conference, UCREL technical paper number 16*. UCREL, Lancaster University.

Storjohann, P. 2005a. 'Corpus-driven vs. corpus-based approach to the study of relational patters' in *Proceedings of the Corpus Linguistics Conference 2005 in Birmingham. Vol. 1, no. 1.* (www.corpus.bham.ac.uk/pclc/).

Storjohann, P. 2005b. '*elexiko* – A Corpus-Based Monolingual German Dictionary.' *Hermes, Journal of Linguistics* 34. Århus, 55-83.

Storrer, A. 2001. 'Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie' in Lemberg, I., Schröder, B., and Storrer, A. (eds.), *Chancen und Perspektiven computergestützter Lexikographie* (Lexicographica Series Maior 107). Tübingen: Niemeyer, 53-69.

Trap-Jensen, L. 2002. 'Descriptive and Normative Aspects of Lexicographic Decision-Making: The Boderline Cases' in Braasch, A., and Povlsen, C. (eds.), *Proceedings of the Tenth EURALEX International Congress*. Copenhagen: CST, 503-508.

The Institute for German Language (Institut für Deutsche Sprache) is the central non-university research institution for the study and documentation of the structure, the contemporary usage, and the recent history of the German language. The IDS was founded in 1964 in Mannheim and is still located there. The research is carried out in three departments, the Department of Grammar, the Department of Pragmatics, and the Department of Lexical Studies. Project clusters on corpus linguistics and research infrastructure, as well as the Central Data Processing Section and the Public Relations Section, complement the work of the institute.

In the Department of Grammar, the grammatical structures of German are identified and described, also including their comparison to other languages. The Department of Pragmatics carries out research on spoken German and linguistic behaviour in conversation. Within the Department of Lexical Studies (Lexik), three main lines of research are currently pursued, (i) lexicology and lexicography from a cultural-historical perspective, (ii) lexical theory with respect to syntagmatic aspects of lexical items, and (iii) corpus-based Internet-lexicography as described in the article at hand.

www.ids.mannheim.de