The Lexicographic Portal of the IDS: Connecting Heterogeneous Lexicographic Resources by a Consistent Concept of Data Modelling

Carolin Müller-Spitzer
Institut für Deutsche Sprache

The Online-Wortschatz-Informationssystem Deutsch *(OWID; Online Vocabulary Information System German) of the* Institut für Deutsche Sprache *(IDS; German Language Institute) in Mannheim is a lexicographic Internet portal for various electronic dictionary resources that are being compiled at the IDS. It is an explicit goal of OWID, not to present a random collection of unrelated reference works but to build a network of actually related lexicographic products. Hence, the core of the project is the design of an innovative concept of data modelling and structuring. The goal of this granular data modelling is to allow flexible access of each individual lexicographic resource as well as access across diverse dictionary resources. At the same time, fine-grained interconnectedness of all resources should be made possible. Every lexicographic resource within OWID—*elexiko, Neologismenwörterbuch, Wortverbindungen online, Schulddiskurs im ersten Nachkriegsjahrzehnt*—accomplishes this requirement with regard to data modelling and structuring. The paper explains the underlying consistent concept of the data modelling for the overall heterogeneous lexicographical resources. Also it is shown, how the modelling potential has been converting into the Internet presence of OWID.*

1. Project outline

The *Online-Wortschatz-Informationssystem Deutsch* (OWID; Online Vocabulary Information System German) of the Institut für Deutsche Sprache (IDS; German Language Institute) in Mannheim is a lexicographic Internet portal for various electronic dictionary resources that are being compiled at the IDS. Originally, it has emerged from the *elexiko* project, which develops a new corpus-based dictionary of contemporary German and has so far aimed at building a lexicographic information portal for the IDS. (cf. Klosa et al. 2006) To elucidate the distinctions between the head portal and the *elexiko*-dictionary itself, the portal was renamed OWID (See *www.owid.de*). The main emphasis of OWID is on academic lexicographic resources with focus on contemporary German. Presently, the following dictionaries are included in OWID.

- *elexiko*: This electronic dictionary consists of an index of about 300,000 short entries with information on spelling (new and old spelling) and syllabication, including information about inflection (from canoo.net) for all headwords. Soon, information (e.g., on word formation) and corpus samples will be added for approximately 250,000 entries for low-frequency lexemes. Furthermore, *elexiko* comprises more than 830 fully elaborated entries of high-frequency headwords, focussing on extensive semantic-pragmatic descriptions of lexical items in actual language use. The primary and exclusive basis for lexicographic interpretation is a comprehensive German corpus, the *elexiko*-corpus. (cf. Storjohann 2005). The dictionary is being extended continuously by further elaborated entries.[1]
- *Neologismenwörterbuch* (Dictionary of Neologisms): This electronic dictionary describes in detail about 800 new words and new meanings of established words, added

[1] For more information on *elexiko*, see Haß (2005) and the list of other project-related publications on *http://hypermedia.ids-mannheim.de/elexiko/ModulElex/pgProjektveroeffentlichungen.html*.

to the German vocabulary during the 1990s.[2] This dictionary is also constantly upgraded.

- *Wortverbindungen online* (Collocations Online): This resource of OWID publishes the research results of the project *Usuelle Wortverbindungen* (fixed multiword combinations), suitable for online presentation. Twenty-five detailed entries for fixed multiword combinations and 100 shorter entries dealing with additional collocations are currently available.[3]
- *Schulddiskurs im ersten Nachkriegsjahrzehnt* (Discourse on Guilt in the First Post War Decade): This dictionary is a reference work corresponding to the study of lexemes that establish the notional area of "guilt" in the early post-war era (1945-55), published in 2005 (Kämper 2005 and 2007). It subsumes lexical-semantic results of this study as dictionary entries, describing various lexical items that constituted this discourse during that time period according to lexicographic principles.[4]

In the near future, the *Handbuch Deutscher Kommunikationsverben* (Harras et al. 2004) (Handbook of German Communication Verbs) with approximately 350 paradigms of communication verbs as well as the VALBU (*Valenzwörterbuch deutscher Verben*, Schumacher et al. 2004) (Valency Dictionary of German Verbs) will be published in OWID.

Even though these lexicographic resources might appear to be very diverse at first glance, it has to be stressed that it has always been an explicit goal of OWID not to present a random collection of unrelated dictionary resources but to build a network of interrelated lexicographic products. Therefore, it was necessary to maintain the independence of each individual dictionary project and, at the same time, to ensure the integration of all these different data. Hence, the core of the project is the design of an innovative concept of data modelling and structuring.

2. Core data modelling

Concerning 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, see also Engelberg and Lemnitzer 2001, and Storrer 2001. For more detailed information on the modelling concept, see Müller-Spitzer 2006, 2007a, and 2007b.) 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 should also be kept open for the possible integration of externally developed lexicographic resources, namely reference works that are written outside the IDS.

The approach chosen here not only guaranteed to connect different lexicographic products beneath the management of OWID on the macro structure level—which means the level of the headwords—but also made it possible to access the dictionaries on a more granular level. Therefore, the attempt was to harmonize modelling on the level of the content structure, that is, the level of the individual lexicographic information unit rather than organizing the different lexicographic processes independently. This is the major reason why OWID, in the

[2] See also Herberg et al. (2004) and *http://hypermedia.ids-mannheim.de/elexiko/ModulNeo/*.

[3] See *http://hypermedia.ids-mannheim.de/elexiko/ModulMV/index.html*.

[4] See *http://hypermedia.ids-mannheim.de/elexiko/ModulSchulddisk/Start.html*.

long run, is meant to be a different type of lexicographic information portal than, for example, the dictionary portal of the Berlin-Brandenburgischen Akademie der Wissenschaften (Berlin-Brandenburg Academy of Science) and of the Heidelberger Akademie der Wissenschaften (Heidelberg Academy of Science) (*www.woerterbuch-portal.de/*), which mainly provide a joint entry page for diverse dictionaries. Generally, a joint entry page for heterogeneous lexicographic products is very useful, but since currently only IDS-internal projects are involved in OWID, the chance should be taken to interlink them on deeper levels and, therefore, to allow for more flexible access possibilities.

OWID uses a single modelling process for all projects: For each individual resource, a specially-tailored XML-DTD and XML-schema was developed. 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 in a XML editor and stored in an Oracle database system. For presentation purposes, 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 which information is accessible across the different dictionaries. This procedure requires each individual information unit to be granularly tagged in all entry structures but also allows for automatic access to each content unit. The following XML detail of the entry "Fon" from the dictionary of neologisms, illustrating the tagging of information on word formation, provides an initial impression of the overall granularity of tagging.

```
<wortbildung><nm-wortbildung>

     <kurzwortbldg>
       <unisegmental typ="endwort">
      <kurzwortA
             artikel-refid="105886" lesart-refid="0"
              ltspez-refid="0"basistyp="nomen">
             Telefon
       </kurzwortA>
       </unisegmental>

     <angabe-zusatz><kommentar><lex-interpretationK>
       <k-absatz>analog zu <obj-text>Fax</obj-text> gebildet</k-absatz>
       </lex-interpretationK></kommentar></angabe-zusatz>

     </kurzwortbldg>

</wortbildung>
```

Within a much larger context, the corresponding XML modelling is only briefly sketched here, and it will be elucidated in much more detail in the presentation, including the illustration of underlying modelling through further examples. (See also Klosa and Müller-Spitzer 2007) This type of data modelling—a singular specially-tailored but explicitly synchronized modelling for diverse lexicographic resources—can be considered to be an innovative approach of a new kind, like Schlaps 2007 and Kunze and Lemnitzer 2007, 85ff. have recently explained.

3. Current state and future perspectives

The goal of this granular data modelling is to allow flexible access of each individual lexicographic resource as well as access across diverse dictionary resources. At the same time, fine-grained interconnectedness of all resources should be made possible. Every lexicographic resource within OWID accomplishes this requirement with regard to data modelling and structuring.

However, the Internet presence of OWID currently shows only first beginnings of its overall potential. An actual benefit is to display information through style sheets flexibly. Thus, not every available information unit is shown by default: Some of them are only used for search; others are only accessible for internal use by members of the project team. It also needs to be stressed that identically modelled data are displayed differently to certain extents. For example, *elexiko* and the neologism dictionary choose quite different methods of layout although most of

the data are modelled equally. Moreover, the search site “Erweiterte Suche” (Advanced Search) of each dictionary enables detailed investigations of specific information units. It is, for instance, possible to search in *elexiko* for all nouns with an old spelling variant that are compounds. Or in the neologismdictionary, for example, you can search for all new lexemes (Neologismentyp=“Neulexem”) that entered the German language in the early 1990s (Aufkommen=“Anfang der 90er Jahre”). (See Fig. 1) Search results are words like “wegzappen”, “Neufünfland”, or “abspacen”. In the same way, it is possible to search for all verbs (Wortart=“Verb”) that gained a new sense (Neologismentyp=“Neubedeutung”) in the 1990s. This question leads the user to 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.

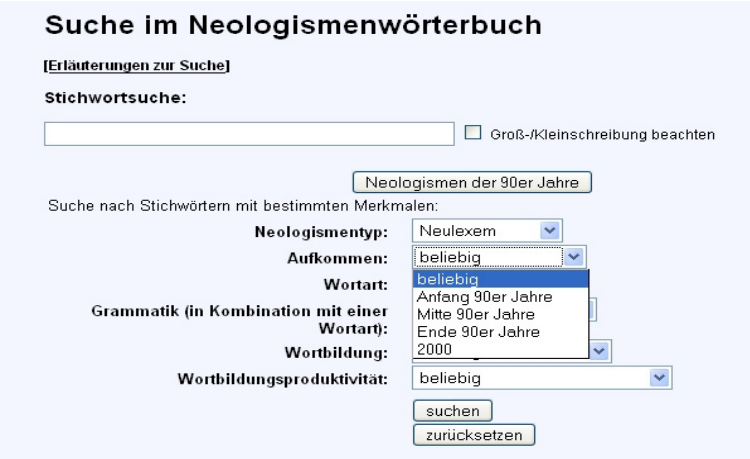

Figure 1. Search window of the neologism dictionary

In the near future, a further step will be taken to provide searches across the different OWID-products. Furthermore, it is our goal to provide maximum flexibility for the user interface of OWID and of the individual dictionaries, respectively. Therefore, the modelling relates solely to the content—clearness of presentation and user aspects. Thus, the same data can be displayed differently for numerous user types and look-up situations with no need to transform the data. But this approach too has its limitations. It would be very fascinating to find out, for example, in which cases it would be necessary to create information units twice with regard to the needs of different user groups. But until now, for electronic lexicography, it has not been systematically investigated which functionalities are explicitly useful for certain user groups and situations. Therefore, we plan an academic project focused on user research as well as on ways of connecting lexicographic resources. So for OWID, today’s challenge is—besides the continuous extension and enhancement of the individual dictionaries and integration of new dictionaries—to focus on providing 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.

References

De Schryver, G. M. (2003). "Lexicographer's Dreams in the Electronic-Dictionary Age". *International Journal of Lexicography* 16 (2). 143-199.

Engelberg, S.; Lemnitzer, L. (2001). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Narr.

Haß, U. (ed.) (2005). Grundfragen der elektronischen Lexikographie. elexiko - das Online-Informationssystem zum deutschen Wortschatz. Berlin: de Gruyter.

Harras, G. et al. (2004). Handbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch. Berlin: de Gruyter.

Herberg, D.; Kinne, M.; Steffens, D. (2004). *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen*. Unter Mitarbeit von Elke Tellenbach und Doris al-Wadi. Berlin: de Gruyter.

Kämper, H. (2005). Der Schulddiskurs in der frühen Nachkriegszeit. Ein Beitrag zur Geschichte des sprachlichen Umbruchs nach 1945. Berlin: de Gruyter.

Kämper, H. (2007). Opfer - Täter - Nichttäter. Ein Wörterbuch zum Schulddiskurs 1945-1955. Berlin: de Gruyter.

Klosa, A.; Schnörch, U.; Storjohann, P. (2006). "ELEXIKO - A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache, Mannheim". In Corino, E.; Marello, C.; Onesti, C. (eds.). *Atti del XII Congresso Internazionale di Lessicografia. Torino, 6-9 settembre 2006 (Proceedings of the 12th EURALEX International Congress)*. Vol. 1. Alessandria: Edizioni dell'Orso. 425-430.

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

Kunze, C.; Lemnitzer, L. (2007). *Computerlexikographie. Eine Einführung*. Tübingen: Narr.

Müller-Spitzer, C. (2007a). "Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung". *Hermes* 38. 137-171.

Müller-Spitzer, C. (2007b). Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis. Tübingen: Narr.

Müller-Spitzer, C. (2006). "Das Konzept der Inhaltsstruktur. Eine Auseinandersetzung mit dem Konzept der Mikrostrukturen im Kontext der Modellierung einer lexikografischen Datenbasis". *Lexicographica* 22. 72-94.

Schlaps, C. (2007). "Grundfragen der elektronischen Lexikographie. *Elexiko* – das Online-Informationssystem zum deutschen Wortschatz. Hg. v. Ulrike Hass. Berlin, New York: de Gruyter 2005. Short review". *Lexicographica* 22. 311-314.

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

Storrer, A. (2001). "Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie". In Lemberg, I.; Schröder, B.; Storrer, A. (eds.). *Chancen und Perspektiven computergestützter Lexikographie*. Tübingen: Niemeyer. 53-69.

Storjohann, P. (2005). "Elexiko - A Corpus-Based Monolingual German Dictionary". *Hermes* 34. 55-83.