First ideas of user-adapted views of lexicographic data exemplified on OWID and *elexiko*

Carolin Müller-Spitzer
Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
mueller-spitzer@ids-mannheim.de

Christine Möhrs
Institut für Deutsche Sprache
R 5, 6-13
D-68161 Mannheim
moehrs@lexik.ids-mannheim.de

Abstract

This paper is a project report of the lexicographic Internet portal OWID, an Online Vocabulary Information System of German which is being built at the Institute of German Language in Mannheim (IDS). Overall, the contents of the portal and its technical approaches will be presented. The lexical database is structured in a granular way which allows to extend possible search options for lexicographers. Against the background of current research on using electronic dictionaries, the project OWID is also working on first ideas of user-adapted access and user-adapted views of the lexicographic data. Due to the fact that the portal OWID comprises dictionaries which are available online it is possible to change the design and functions of the website easily (in comparison to printed dictionaries). Ideas of implementing user-adapted views of the lexicographic data will be demonstrated by using an example taken from one of the dictionaries of the portal, namely *elexiko*.

1 Project report

The *Online-Wortschatz-Informationssystem Deutsch* (OWID; Online Vocabulary Information System of German), a project of the *Institut für Deutsche Sprache* (IDS; Institute of German Language) in Mannheim is a lexicographic Internet portal containing both, various electronic dictionary resources that are currently being compiled at the IDS on the one hand and external resources on the other hand which will be included additionally in the near future (cf. www.owid.de). Originally, the project had its roots based in the IDS project *elexiko*, a lexicographic enterprise, which develops a new corpus-based dictionary of contemporary German. It formed the basis of a lexicographic information portal for the IDS (cf. Klosa et al. 2006). The main emphasis of OWID is on the integration of different academic lexicographic resources with the 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 and syllabication, including information about inflection (from www.canoo.net). In the near future, further information (e.g. on word formation) and corpus samples will be added for all lexemes. Furthermore, *elexiko* comprises over 900 fully elaborated entries of headwords which are highly frequent in the underlying corpus. These contain extensive semantic-pragmatic descriptions of lexical items in actual language use. The dictionary is being extended continuously by further elaborated entries (cf. Klosa et al. 2006).
- *Neologismenwörterbuch* (Dictionary of Neologisms): This electronic dictionary describes about 800 new words and new meanings of established words in detail which emerged in the German vocabulary during the 1990s. This dictionary is also being upgraded constantly.
- *Wortverbindungen online* (Collocations Online): This resource of OWID publishes the research results of the project *Usuelle*

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 39–46
Manchester, August 2008

Wortverbindungen. These concern different fixed multiword combinations. Currently, 25 detailed entries for fixed multiword combinations and 100 shorter entries dealing with collocations are available to users.

- *Diskurswörterbuch 1945-55* (Discourse Dictionary 1945-55): This dictionary is a reference work resulting from a larger study of lexemes that establish the notional area of "guilt" in the early post-war era (1945-55), published in 2005.

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

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 while, at the same time, to ensure the integration of all the different data. Even though, the different lexicographic resources may appear to be very diverse at first glance, they share some of their data modelling features. Both, the common intergration and the individual independence of each project are reflected in the current online presentation of the portal. On the welcome page of OWID the user can choose which dictionary s/he wants to use. If s/he looks up a word in all dictionaries of the portal there is a coloured marker indicating the corresponding dictionary resource (black = *elexiko*, blue = Neologism, green = Discourse dictionary, red = Collocations). In addition, there are links and cross-references between the products (see for example the interrelation between the entry "Liebe macht blind" in the dictionary "Collocations Online" and the entries "Liebe" / "blind" in *elexiko*). This kind of interrelation will be expanded in the future.

Another goal is to provide a basis for user-adapted access to the lexicographic data. "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). Hence, the core of the project is the design of an innovative concept of data modelling and structuring.

2 Data Modelling

As emphasised before, the contents of 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 were to be developed and if the lexicographic data had been

OWID DTD-library				
modules for the whole OWID portal		allg-entities.dtd (DTD for general entities)	allg-elemente.dtd (DTD for general elements and attributes)	
modules for cross-dictionary object groups	ewl-objekte.dtd (DTD for objects of single-word-items)	mwl-objekte.dtd (DTD for objects of multi-word-items)	ewl_mwl-objekte.dtd (DTD for objects of single-word-and multi-word-tems)	ewl-grammatik.dtd (DTD for grammatical objects)
modules for object groups of specific dictionaries		elexiko-allgobj.dtd (DTD for general objects of *elexiko*)	neo-allgobj.dtd (DTD for general objects of the neologism-dictionary)	
Head-DTDs for each dictionary	elexiko-ewl.dtd (Head-DTD for *elexiko*)	neo-ewl.dtd (Head-DTD for single-word-items of the neologism-dictionary) neo-mwl.dtd (Head-DTD for multi-word-items of the neologism-dictionary)	mwl.dtd (Head-DTD for multi-word-items of the project "Usuelle Wortverbindungen")	zeitreflexion1945-55.dtd (Head-DTD for the discourse-dictionary 1945-55)

Table 1. OWID DTD-library

interlinked even more adequately. So on the one hand, in order to guarantee a basis for a common access structure to the all contents, consistent principles for modelling and structuring the contents were applied to all integrated products. On the other hand, OWID is also kept open for the possible integration of externally developed lexicographic resources, namely reference works that are written outside the IDS. However, externally compiled data has to be structured in accordance to the OWID modelling concept.

The approach chosen here not only guarantees to connect different lexicographic products under the management of OWID on the macro structure level – which means the level of the headwords – but also makes it possible to access the dictionaries on a more granular level. OWID attempts to harmonise 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.

OWID uses a single modelling process for all projects: For each individual resource, a specifically-tailored XML-DTD and XML-schema were developed respectively. Each individual information unit is granularly tagged in all entry structures, so that automatic access to each content unit is ensured. The dictionary entries are then written in an XML editor and stored in an Oracle database system. For presentation purposes, the XML data are transformed by an XSLT stylesheet to HTML (cf. Müller-Spitzer 2007).

A DTD library was created for OWID where specific DTDs contain all entities, elements, or attributes that are shared by all entry structures in order to provide a uniform structure for lexicographic information of the same type which is contained in the different dictionaries (cf. Tab. 1). The modelling shows which information is accessible across the different dictionaries (the results from the different dictionaries are marked in different colours). This type of data modelling – a singular specificially-tailored but explicitly synchronised modelling for diverse lexicographic resources – can be considered to be an innovative approach of a new kind, as Schlaps (2007) and Kunze / Lemnitzer (2007) have recently explained.

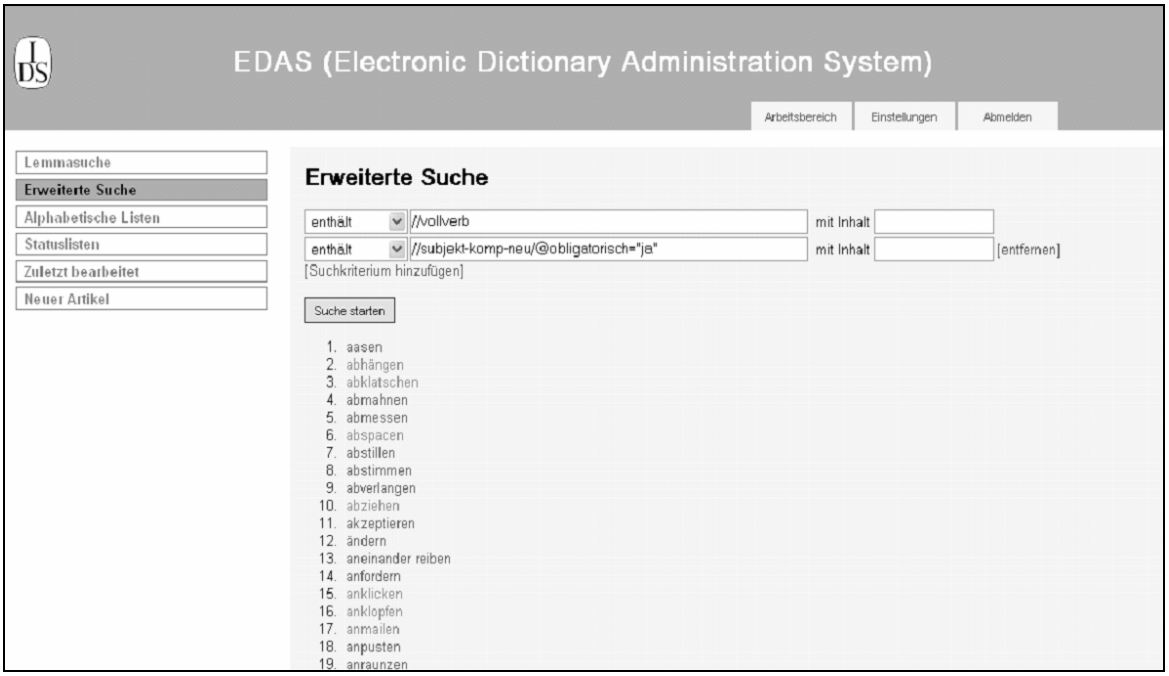

Figure 1: Advanced search options for lexicographers

We decided to use a specifically-tailored modelling because the XML-structure also serves as a model for compiling the lexicographic entries in the XML-Editor. What this means for lexicographers is that the more individually customised the XML-structure is, the less one needs an additional manual for comply with the entry structure. However, one could easily transform this structure into a specific standard such as LMF or TEI because the structure is very fine-grained. The following XML detail of the entry “emailen” from the Dictionary of Neologisms illustrating the tagging of information on valency gives an example for the overall granularity of tagging.

```
   <vb-valenz-neu>

   <satzbauplan>
   <satzbauplanA>jemand emailt (jemandem) (et-
was)</satzbauplanA>

   </satzbauplan>
   <satzbauplan>
   <satzbauplanA> jemand emailt (etwas) an je-
manden</satzbauplanA>
   </satzbauplan>

   <satzbauplan>
   <satzbauplanA>jemand      emailt,      dass
[...]</satzbauplanA>
   </satzbauplan>

   <vb-komplemente-neu>

   <subjekt-komp-neu obligatorisch="ja">
   <nom-nominalphrase-neu/>
   </subjekt-komp-neu>

   <objekt-komp-vb obligatorisch="nein">
   <dat-nominalphrase-vb/>
   </objekt-komp-vb>

   <objekt-komp-vb obligatorisch="nein">
   <akk-nominalphrase-vb/>
   <dass-satz-vb/>
   </objekt-komp-vb>

   <objekt-komp-vb obligatorisch="ja">
   <praepositionalphrase-vb        praeposi-
on="an"/>
   </objekt-komp-vb>

   </vb-komplemente-neu>
   </vb-valenz-neu>
```

Within our internal editorial system, lexicographers are able to use this structure for advanced searches (with XPath expressions). For example, one can search for all regular verbs `(//vollverb)` which have obligatory object complements `(//objekt-komp-vb/@obligatorisch="ja"` which are realised as a dative NP `(//dat-nominalphrase-vb)`. In this example, the search results are entries from *elexiko* as well as from the neologism-dictionary (cf. Fig. 1). We are planning to provide these extended search options also for users.[2]

Moreover, it would be possible to involve the user in the process of deciding which information should be presented on the website. As explained, every information unit in the dictionaries is encoded separately. Against this background, we can think of customizing the microstructure by the users themselves (in addition to the extended search for example in *elexiko*). So the user could select the type of information s/he wants to use individually. Fig. 2 shows what such a presentation could look like. At the top of the page, the user is able to select the type of information which s/he wants to see directly underneath. If s/he wants to change the options s/he can use the update button in order to modulate the desktop view. In this example, the two different senses of the entry 'Meer' are shown side by side with the chosen kind of information (here the definition together with typical uses of the headword). This kind of presentation enables the users to compare this information given for the two senses at one sight.

3 Research on using electronic dictionaries

Research on using dictionaries is a core field of study in lexicography (cf. Wang 2001 or Atkins 1998). Fortunately, in the last two decades, research on using printed dictionaries has attracted the attention of more researchers. Although Engelberg and Lemnitzer had noticed in 2001 that there are only little inquiries about influences on the users' behaviour in relation to innovations in the field of electronic lexicography (cf. Engelberg and Lemnitzer 2001), in the last few years research on electronic dictionaries has grown.

Such metalexicographic research plays a major role with regard to monitoring the dictionary user on the Internet – for example in the analysis of log-files. At the moment, there are not many research reports about the analysis of log-files. "Although the proposal to draw upon log files in order to improve dictionaries was already expressed in the mid-1980s […], and although numerous researchers have reiterated this idea in recent years […], very few reports have been published of *real-world* dictionaries actually marking use of this strategy" (de Schryver and Joffe 2004, 187). The studies and methods mentioned here are interesting for research on using electronic dictionaries especially because an electronic dictionary is a product which can be modulated and updated immediately. Log-files can show what the user has inserted into the search box and how the user has navigated (cf. de Schryver and Joffe 2004). However, good results are only seen with this method if the database of the dictionary is created with a flat structure. In the actual log- files we only see which word the user has typed in the search box. We can not easily detect in which way and how comfortly the user navigates through the entry or

[2] The development of the *Electronic Dictionary Administration System* (cf. Fig. 1) is a work of Roman Schneider, a researcher of the IDS.

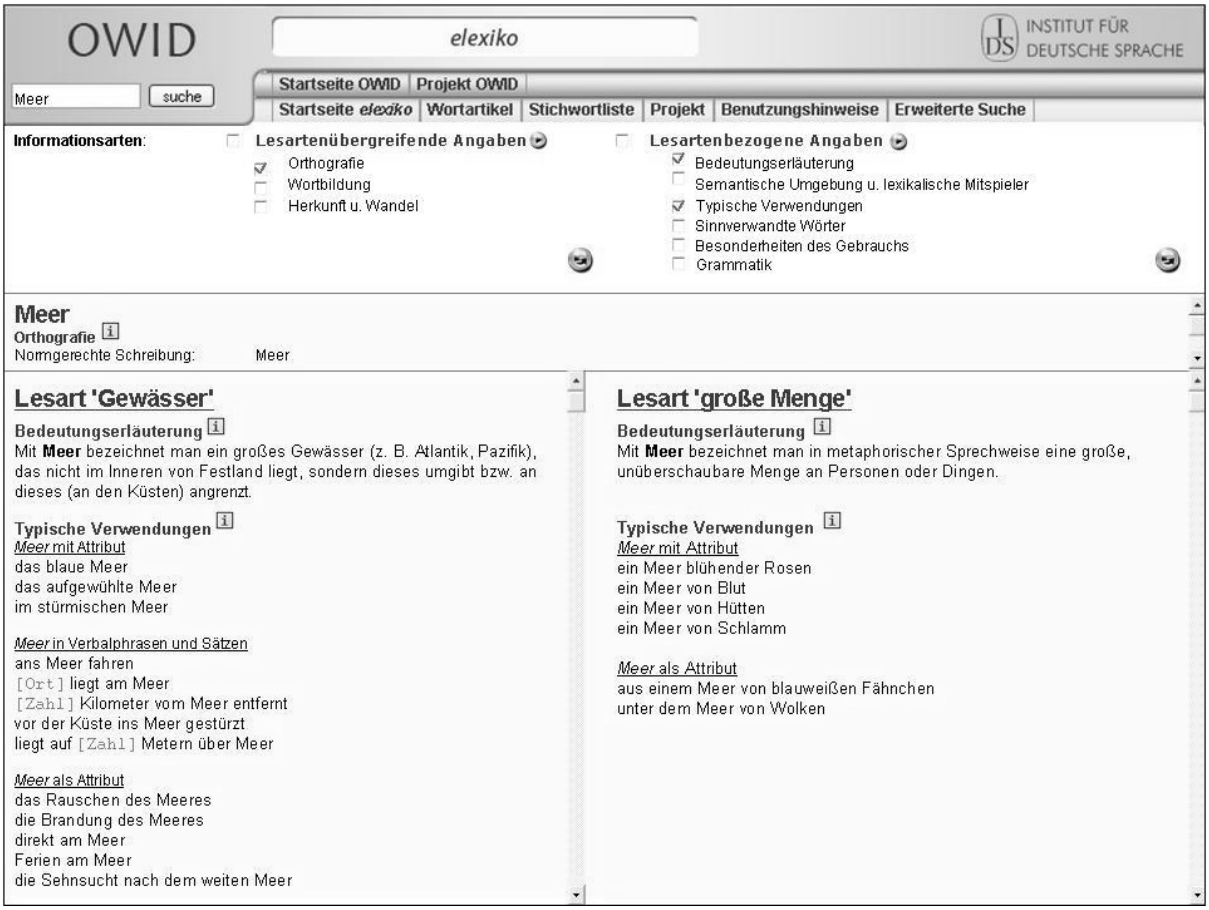

Figure 2. Online view of *elexiko* with an information display for customizing the microstructure dynamically

which information s/he has looked at more closely. However, this is exactly the type of information we are looking for. Therefore, other methods like standardised evaluation, interviews etc. also have to be taken into account. Analysing log-files can not substitute these methods alone.

OWID is also gradually putting user research into practice: Firstly, OWID has been making use of the analysis of log-files for some time. Secondly, a standardised online survey was conducted in the context of an MA thesis (cf. Scherer 2008). Finally, a short study based on interviews of OWID and in particular of *elexiko*, one of the dictionaries of the portal, was carried out.

Although currently the modelling is used mainly in the lexicographic process there is still a lot of room for further development of the abilities to present the structured information. The capability of data modelling in OWID should be visible for lexicographers as well as for users (cf. Müller-Spitzer 2007). Involving the user and his/her requirements in searching and navigating through OWID is the starting point for defining user-adapted views of the lexicographic data.

4 Defining user-adapted Views

As shown above, the lexicographic contents are structured granulary and strictly content-based. This technology allows to define user-adapted views of the lexicographic data. Printed dictionaries cannot offer this option. A printed dictionary is designed for a specific user type and for specific situations of use as a whole. In OWID, the data for electronic dictionaries is initially organised independently of its users. In a second step, lexicographic information can be used as the foundation of the definition of user-specific layers (e.g. based on the technology of XSLT-stylesheets) in order to filter relevant data for a specific situation of use “on demand”. Knowlegde on what users prototypically look for in printed dictionaries is established by numerous research works. For example someone who uses a dictionary to understand a text wants to get a short overview on the meaning of a word. If someone has to produce a text it is more helpful to get word information about correct spelling, grammar, typical uses, collocations or sense-related items. Furthermore lexicographers of

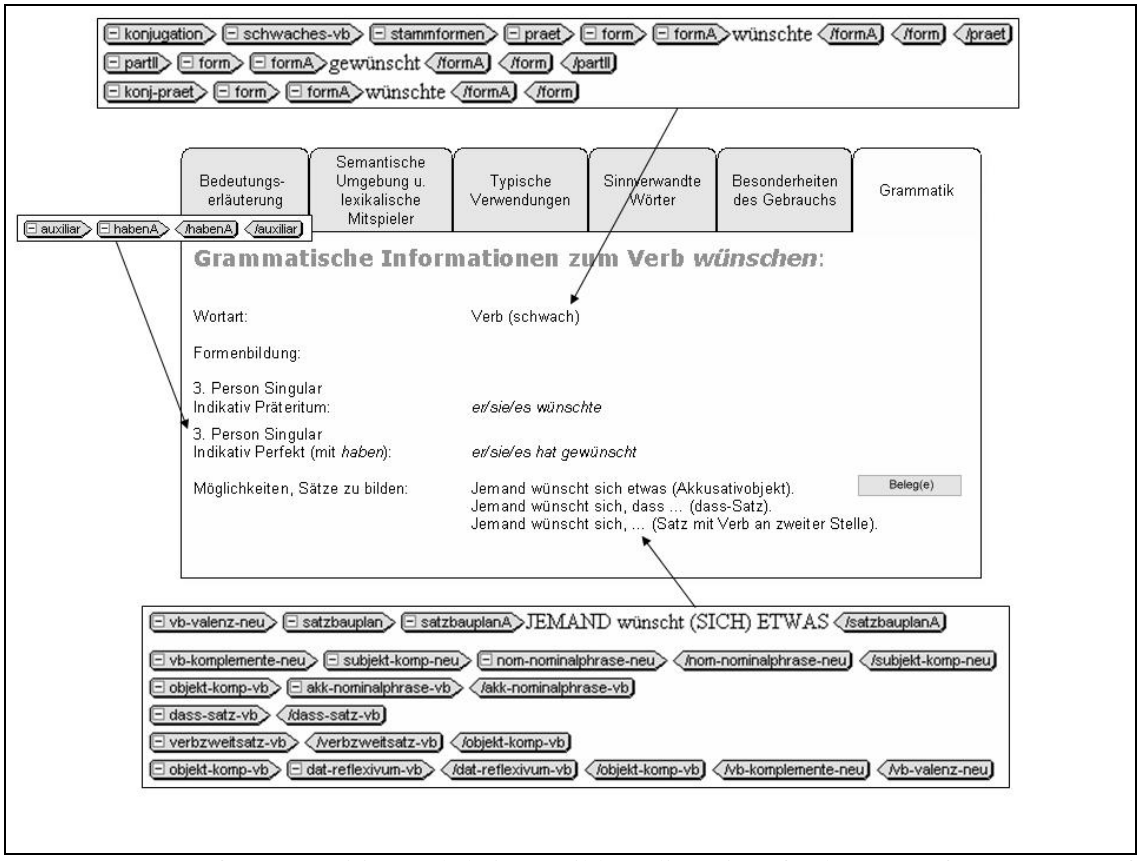

Figure 3. Extracts of XML-entities and their possible online view for learners of German as a foreign language (entry **wünschen**, part "Grammar" for the meaning 'ersehnen')

electronic dictionaries can go into detail about the demands of learners of German as foreign language (L2-Learners) resp. German native speakers. By taking this into consideration, one can think of developing different profiles for different user situations. According to a chosen profile the lexicographic information is then presented in a specialised way. This would be another form of a user-adapted view (besides customizing the microstructure dynamically as it is shown in Fig. 2). In *elexiko*, one of the dictionaries of OWID, the online view presents the lexicographic data in one standardised view. However, the technical conditions can also allow to show the same XML-data of an entry in different ways for different user groups. As an example one can see the part "Grammar" in *elexiko* in Figure 3 and 4 differing in comprehensiveness. Detailed information on inflection and word order are very important for L2-Learners. Therefore such information is presented more extensively in Fig. 3. In comparison native speakers know intuitionally the inflection of words or the realization of different sentence constructions. In Fig. 4 one can see a shortened presentation of grammatical information of the same XML-data. This example illustrates the general principle of defining different user-adapted views of one lexicographic data. It is important that the different user-adapted presentations of the part "Grammar" in *elexiko* or every other part of word information in *elexiko* can be realised without changing the data. The only change happens in the stylesheet. Other views completely different from the actually used stylesheet can be imagined easily. We will discuss further examples in the talk.

For a printed dictionary it is sufficient to define the types of information that shall be included for the intendend user. Questions of presentation are discussed on this basis and along the strong tradition for the layout of printed dictionaries. When compiling user-adapted views of a general lexicographic data for an electronic medium we have to consider:

How do users navigate in electronic dictionaries especially in a dictionary portal? How do they use the search options? Which form of nesting the specific word information is user friendly and when does clearness suffer? (Cf. Almind 2005) More specifically we need to ask: Should a user (i.e. while using a dictionary) create a profile at

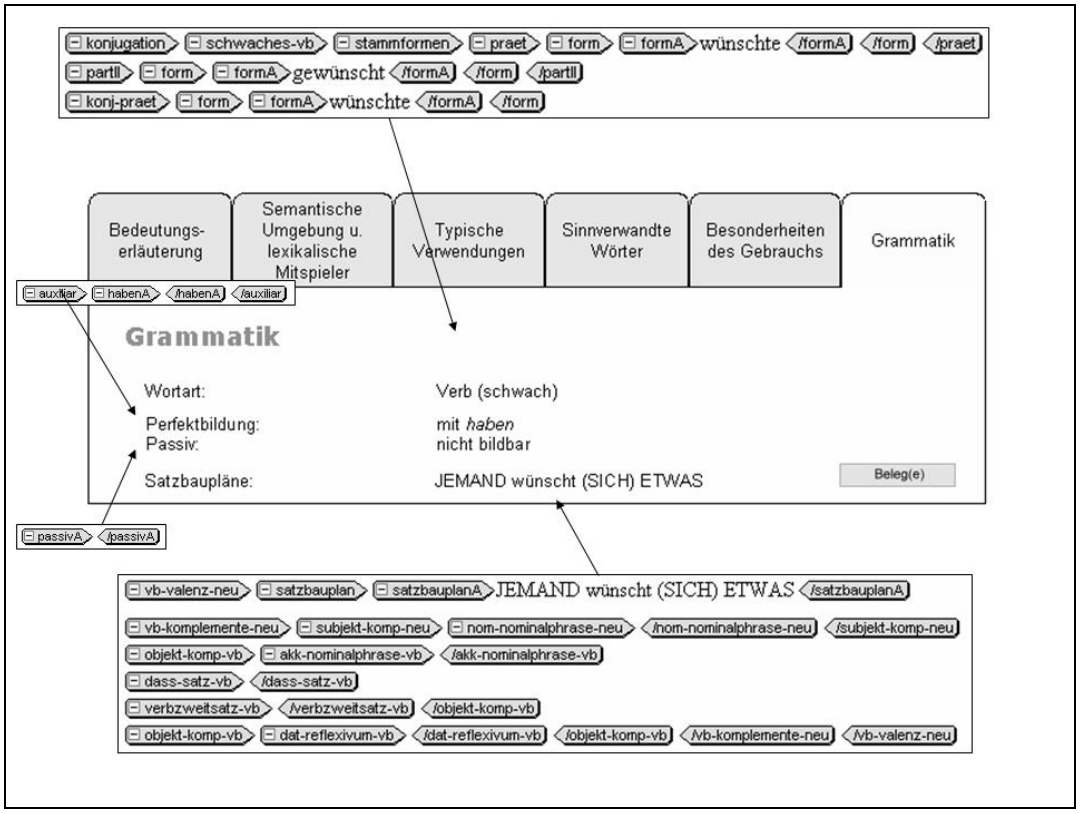

Figure 4. Extracts of XML-entities and their possible online view for German native speakers

the beginning of a session (e.g. user type: non-native speaker, situation of use: reception of a text) and should s/he navigate in all articles with this profile? Or is it more user friendly to being able to change ones profile and look at the same entry with different profiles which means customizing the microstructure dynamically?

As OWID fulfills all technical requirements for a user-adapted presentation, as shown above, this project will be able to realise innovative forms of access to the lexicographic data. Research on the use of the dictionaries published in OWID will be the basis on which different forms of presentation will be developed.

References

Almind, Richard. 2005. *Designing Internet Dictionaries*, in: Hermes 34:37-54.

Atkins, B. T. Sue (Ed.) (1998): *Using dictonaries. Studies of dictionary use by language learners and translators.* (= Lexicographica. Series maior 88), Tübingen.

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

De Schryver, Gilles Maurice / Joffe, David. 2004. *On How Electronic Dictionaries are Really Used*, in: Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France. Vol. I, ed. by Geoffrey Williams / Sandra Vessier:187–196.

Engelberg, Stefan / Lemnitzer, Lothar. 2001. *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Klosa, Annette / Schnörch, Ulrich / Storjohann, Petra. 2006. *ELEXIKO - A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache*, Mannheim, in: Proceedings of the 12th EURALEX International Congress (Atti del XII Congresso Internazionale di Lessicografia), EURALEX 2006, Turin, Italy, September 6th-9th, 2006. Vol. 1, ed. by Carla Marello et al., Alessandria:425-430.

Kunze, Claudia / Lemnitzer, Lothar. 2007. *Computerlexikographie. Eine Einführung*. Tübingen: Narr.

Müller-Spitzer, Carolin (2007): *Das elexiko-Portal: Ein neuer Zugang zu lexikografischen Arbeiten am Institut für Deutsche Sprache*, in: Datenstrukturen für linguistische Ressourcen und ihre Anwendungen. Proceedings of the Biennial GLDV Conference 2007 (April 11-13, 2007, Eberhard Karls Universität Tübingen), ed. by Georg Rehm / Andreas Witt / Lothar Lemnitzer:179-188.

Scherer, Tanja. 2008. *Umsetzung von Zugriffsstrukturen bei Online-Wörterbüchern.* Unveröffentlichte Magisterarbeit an der Universität Mannheim, Phi-

losophische Fakultät, Seminar für Deutsche Philologie, Germanistische Linguistik (Prof. Dr. L. M. Eichinger).

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

Wang, Weiwei. 2001. *Zweisprachige Fachlexikographie. Benutzungsforschung, Typologie und mikrostrukturelle Konzeption*, Frankfurt a.M. (= Angewandte Sprachwissenschaft 8).