Refining and Exploiting the Structural Markup of the eWDG

Thomas Schmidt
Universität Hamburg

Alexander Geyken
Berlin-Brandenburgische Akademie der Wissenschaften

Angelika Storrer
Universität Dortmund

In this paper, we describe a semi-automated approach to refine the dictionary-entry structure of the digital version of the Wörterbuch der deutschen Gegenwartssprache *(WDG, en.: Dictionary of Present-day German), a dictionary compiled and published between 1952 and 1977 by the Deutsche Akademie der Wissenschaften that comprises six volumes with over 4,500 pages containing more than 120,000 headwords. We discuss the benefits of such a refinement in the context of the dictionary project* Digitales Wörterbuch der deutschen Sprache *(DWDS, en: Digital Dictionary of the German language). In the current phase of the DWDS project, we aim to integrate multiple dictionary and corpus resources in German language into a digital lexical system (DLS). In this context, we plan to expand the current DWDS interface with several special purpose components, which are adaptive in the sense that they offer specialized data views and search mechanisms for different dictionary functions-e.g. text comprehension, text production-and different user groups-e.g. journalists, translators, linguistic researchers, computational linguists. One prerequisite for generating such data views is the selective access to the lexical items in the article structure of the dictionaries which are the object of study. For this purpose, the representation of the eWDG has to be refined. The focus of this paper is on the semi-automated approach used to transform eWDG into a refined version in which the main structural units can be explicitly accessed. We will show how this refinement opens new and flexible ways of visualizing and querying the lexicographic content of the refined version in the context of the DLS project.*

1. Introduction

In this paper, we describe a semi-automated approach to refine the dictionary entry structure of the digital version of a six-volume dictionary of German, and we discuss the benefits of such a refinement in the context of the DWDS project.

The DWDS project, which we briefly describe in section 2, integrates multiple dictionary and corpus resources on the German language in a Digital Lexical Information System. Our paper focuses on one of the dictionary resources in this system, namely on the "Wörterbuch der deutschen Gegenwartssprache" (abbreviated WDG). Section 3 describes this resource and sketches how the paper dictionary was transformed into a dictionary database (eWDG.1) and published online in the first stage of the project. The Lexical Information System is planned as an adaptive system which can be customized for different usage scenarios and user groups. One prerequisite for this adaptability is the selective access to the lexical items in the article structure of the dictionary database. For this purpose, we developed a semi-automated approach to transform eWDG.1 into a refined version (henceforth eWDG.2). The main focus of the paper will be on this transformation, which is described in section 4. In section 5, we show how this structural refinement opens new and flexible ways of visualizing and querying the lexicographic content of the eWDG.2 in the framework of our lexical information system.

2. Project Context: Towards a Lexical Information System (LIS)

The dictionary project "Digitales Wörterbuch der deutschen Sprache" (DWDS, "Digital Dictionary of the German Language") was launched in 2000 at the Berlin-Brandenburg Academy of Sciences (BBAW) with the goal to build an online lexical information system that provides access to multiple German text corpora and digital dictionaries.

In the first phase of the project (2000 to 2007: cf. Klein and Geyken 2000, Klein 2004a, 2004b), we developed an information system in which four different types of resources can be consulted online[1] (Geyken 2005):

- The *dictionary database* eWDG.1, which is described in more detail in section 3.
- The *corpus component* (currently 800 Mio tokens in total) comprises newspaper corpora, specialized corpora (e.g. spoken language, language of the former German Democratic Republic GDR), and the DWDS core corpus, a balanced corpus of German texts from the 20th century. The core corpus consists of 100 million tokens (comparable in size to the British National Corpus), equally distributed over time and over the following five text types: journalism (approx. 27% of the corpus), literary texts (26%), scientific literature (22%), other non-fiction (20%), and transcripts of spoken language (5%). The corpus is encoded according the guidelines of the Text Encoding Initiative (TEI-P5), lemmatized with the TAGH morphology (Geyken and Hanneforth 2006), and tagged with the part-of-speech tagger moot (Jurish 2004) according to the conventions of the Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999). The corpus search engine DDC (Dialing DWDS Concordancer) supports linguistic queries on several annotation levels (word forms, lemmas, STTS part-of-speech categories) and offers filtering (author, title, text type, time intervals) and sorting options (date, sentence length). Details on the design of the corpora and on the technical background of the corpus tools are given in Geyken 2007.
- On the basis of our corpora, the *collocation component* offers several options to compute collocations for a lexical unit according to common statistical measures (mutual information, t-score and log-likelihood).
- An additional *thesaurus component* computes synonyms, hyponymy and hypernyms for lexical units on the basis of the dictionary data (Geyken and Ludwig 2003).

Currently, we are working on the extension of our lexical information system with additional (electronic versions of) print dictionaries and new corpus resources: First, we plan to extend the dictionary component with the electronic edition of the "Deutsches Wörterbuch" (DWB 1964). This dictionary comprises 320,000 entries and will thus increase the number of available dictionary entries to almost 400,000 entries.[2] Second, we will add more texts to the corpus database, in particular texts from the XVIIIth and XIXth centuries: the project "Deutsches Textarchiv"[3] (also situated at the BBAW, funded by the Deutsche Forschungsgemeinschaft) is currently compiling a balanced text corpus with material from the period between 1780 and 1900. It will comprise approximately 50 million tokens and will be made available in 2010.

[1] URL: *www.dwds.de* (approximately 5 million page impressions—PI—per month).

[2] The exact number of entries depends on whether one decides to count only the original entries, or to take into account the derived (compound) entries (cf. section 3). In the lexical intersection of WDG and DWB, entries from both dictionaries will be displayed in the lexical information system. Obviously, in this case, a refined structural mark-up for both dictionaries is necessary so that summaries of the dictionary article(s) can be presented to the user. The refinement of the structural mark-up described in the remainder of this paper is crucial for producing such article summaries.

[3] *www.deutsches-textarchiv.de*.

In this and other work on the lexical information system, we follow a set of general conceptual guidelines (cf. Klein and Geyken 2000, Storrer 2001, Klein 2004a):

a) Dictionary databases and text corpora can be accessed through an *integrated user interface*. With the help of this user interface, a user can supplement a dictionary lookup in the dictionary component with additional information derived from the corpus component. In the simplest case, this additional information consists in corpus examples for a given search term (possibly filtered using parameters like time or text type). Beyond this, the user interface also provides access to automatically computed lexical information, like frequency, co-occurrence and collocations, which "traditional" dictionaries usually do not offer. For this purpose, the project employs robust computational linguistic methods like word sense disambiguation and shallow parsing, thereby continually refining the corpus annotation and the functionality of the query tools.

b) The lexical information system is open for the *incorporation of image, audio and video data*. As a first step in this direction, we are currently integrating sound files illustrating the pronunciation of headwords. The goal is to have such recordings for all main headwords of the WDG dictionary as well as for 30,000 additional high-frequency words from the DWDS corpora. To date (March 2008), more than 60,000 keywords of the WDG have been recorded by a professional speaker. These are currently being evaluated by the Institut für Sprechwissenschaft in Halle. The first 30,000 recordings will be made publicly available in summer 2008.

c) The different dictionary databases are *interlinked*, step-by-step, using semi-automatic methods.

d) The workflow for the treatment of lexicographic entities does not proceed alphabetically, but is organized into *phenomena-related modules*. This means that individual components for specific linguistic levels are completed one-by-one and integrated into the larger system.

   The afore-mentioned pronunciation component, which is currently in the focus of our work, is one example of such a module. Another module, also nearing completion, is concerned with the mapping of headwords to the form prescribed by the German Spelling Reform. The aim is to identify all words affected by the new orthography, to note whether or not the spelling of these words is still valid, and to record their possible variants. Furthermore, a link to the relevant part in the spelling regulations will be added. In cases of doubt or uncertainty (e.g. for the spelling of complex participles as in "alleinstehend" vs. "allein stehend"), we are consulting a member of the Rat für Deutsche Rechtschreibung, Peter Eisenberg. A third module is the idioms database of the Wolfgang-Paul-Preis project (Fellbaum 2004, 2007), which will be integrated into the lexical information system.

   Besides facilitating the organizational management of our project, this modularization of the workflow makes it easier to systematically call on expertise from other research organizations whenever the respective component makes this necessary.

e) The lexical information system aims to be *adaptive* in the sense that we want to offer specialized views on the lexical data for different dictionary functions (e.g. text comprehension, text production, linguistic research), as well as specialized search mechanisms for different user groups (e.g. journalists, translators, linguistic researchers, computational linguists). From the outset, the development of such user-specific views and search mechanisms is tested and evaluated with the respective user groups, and can thus be continually optimized.

Especially guidelines (c) to (e) place specific demands on the digital representation of the dictionary resources: incorporating non-textual data, interlinking dictionaries, adding new modules and customizing views and search options all require a fine-grained structuring of

dictionary articles; and it is the articles' content structure rather than their layout or typesetting properties which has to be represented in order to carry out these tasks.

3. WDG and the dictionary database eWDG

The focus of this paper is on the dictionary database eWDG, which is based on a six-volume paper dictionary, the "Wörterbuch der deutschen Gegenwartssprache" (WDG, "Dictionary of Present-day German") published between 1952 and 1977 and compiled at the Deutsche Akademie der Wissenschaften[4]. It comprises six volumes with over 4,500 pages and contains more than 60,000 headwords (more than 120,000 if compounds are counted separately). The term "deutsche Gegenwartssprache" ("German present-day language") is understood in a broad sense by the lexicographers; the dictionary is not restricted to the language spoken and written in the middle of the XXth century, but also incorporates sources from the XVIIIth and XIXth centuries as far as these are still widely read (cf. Malige-Klappenbach 1986, Wiegand 1990).

In 2002, the BBAW commissioned the "Kompetenzzentrum für elektronische Erschließungs— und Publikationsverfahren in den Geisteswissenschaften" at the University of Trier to produce a digital copy of the WDG. Based on the first edition of the printed WDG, the original digitization was done in China and first corrections were performed in Trier between May and July 2002, resulting in TEI-conformant files. At the BBAW, this version was further annotated and transformed into a dictionary database which has been available online since 2003. Apart from the correction of remaining errors, post-processing of the eWDG at the BBAW has been concerned mainly with the dictionary macro structure: using a semi-automatic method, embedded compounds as well as prefix and suffix derivations were annotated in such a way that they can be directly accessed as headwords by a dictionary user. The number of accessible headwords could thus be increased to 120,000. Moreover, information about synonyms and hypernymy/hyponymy was automatically calculated for about 65,000 entries and can also be accessed via the DWDS system.

4. Refining the structural mark-up of the eWDG

As described in the previous section, version 1 of the eWDG has TEI-conformant, content-based mark-up of the macro-structure of the dictionary (the entries) and of the main structural divisions underneath the entry (grammatical information and different senses). However, as figure 1 illustrates, the mark-up of the textual material underneath the sense elements describes typographic properties rather than content structure. The refinement of the structural mark-up of version 2 of the eWDG consists of a transformation of this typography-oriented mark-up into content-oriented mark-up, as in figure 2. This means that definitions ("def"), different types of examples ("eg") and other content-oriented elements are identified and marked as such.

```
<sense>
 <hi rend="bold">10.</hi>
 <hi rend="spaced">salopp</hi>
 es gibt was ab
 <hi rend="italic">es ist etwas Unangenehmes zu gewärtigen</hi>
 : heute kann, wird es noch (et)was a. (
 <hi rend="italic">regnen, ein Gewitter geben</hi>
 ); jmdm. eins a. (
 <hi rend="italic">jmdm. einen Schlag, Tadel versetzen</hi>
 )
</sense>
```

Figure 1: Typography-oriented mark-up of a sense of the headword "abgeben" ('to give off')

[4] Since 1972: Akademie der Wissenschaften der DDR.

```
<sense level="1" n="10.">
 <!-- usage: 'informal' -->
 <usg>salopp</usg>
 <!-- pattern example: a fixed multi-word expression containing the headword -->
 <eg type="pattern">es gibt was ab</eg>
 <!-- definition: 'something unpleasant is to be expected' -->
 <def>es ist etwas Unangenehmes zu gewärtigen</def>
 <!-- illustrating example: an invented sentence containing the headword -->
 <eg type="illustrating">heute kann, wird es noch (et)was a.
   <!-- a paraphrase explaining the use of the headword in this particular example: -->
   <!-- 'es gibt etwas ab' in this example may mean 'there will be rain, a thunderstorm'
-->
   <seg type="paraphrase">(regnen, ein Gewitter geben)</seg></eg>
 <eg type="illustrating">jmdm. eins a.
    <seg type="paraphrase">(jmdm. einen Schlag, Tadel versetzen)</seg></eg>
</sense>
```

Figure 2: Content-oriented mark-up of a sense of the headword "abgeben" ("to give off")[5]

This transformation of presentational to content-oriented mark-up is a typical step in the digitization of printed dictionaries. As early as 1989, Alshawi, Boguraev and Carter (1989: 41ff), describe a very similar task for the LDOCE dictionary:

> The electronic source of the LDOCE is a tape, containing the original data given by the publisher to the printer. Computer tapes, typically typesetting ones, are the usual medium for distributing dictionaries in machine readable form [...]. [However,] typesetting information on its own does not provide a sufficient "handle" on the problems concomitant with loading a dictionary into a database. [...] The issue here is that of recovering, and appropriately labeling the logical units within the entry.

While our task is thus similar in nature to what was done in the LDOCE project, it greatly differs with respect to the technical details and the technology we have at hand. The Unicode and XML standards provide the general framework in which we carry out the task, i.e. they tell us how to digitally represent individual characters and structural entities. Technologies like XPath (for matching patterns in XML documents), XSLT (for carrying out transformations on XML documents) and JDOM (for reading and manipulating XML document in computer memory) support us in the implementation of the transformation routines. Finally, the TEI guidelines provide a set of well-defined categories which we can use to label the resulting logical units. This is described in more detail in the following section.

4.1. *Target format*

For the mark-up of the target structure, we use the tags defined in the TEI P5 guidelines (most of them from module 9 "Dictionaries", but partly also from modules 3 "Elements available in all dictionaries", 16 "Linking, segmentation and alignment" and 17 "Simple analytic mechanisms"). Our main concern at this stage is to cater for a basic compatibility between the eWDG.2 and other electronic dictionaries as well as the DWDS corpus while keeping the original text intact. A more far-reaching standard compliance (e.g. with ISO 1951), possibly involving a partial rearrangement of the dictionary articles, can be aimed for at a later stage.

The following are the most important tags used:

- `<def>` marks a definition. Typically, there is one definition per sense, but there are also senses which are described by examples only. Some senses contain two or more definitions, the later of which refine or supplement the preceding one(s) (e.g. one sense of "Chanson" is defined first as "weltliches, geselliges Lied"—"wordly, sociable song"—, and then as "Heldenlied"—"heroic song"—). Occasionally, definitions have a narrower scope than the entire sense and only define an idiomatic or figurative use of the headword.

[5] Here and in figures 4 and 5, XML style comments (enclosed in "<!—" and "-->") serve to illustrate the examples for the reader. They are not a part of the actual dictionary data.

- `<eg>` marks an example. We use a `type` attribute to distinguish three different kinds of examples: Quoted examples are citations from external sources, i.e. from the literature or from contemporary journalistic or scientific articles. Illustrating examples are invented by the lexicographer to demonstrate typical usages and collocates of the headword. Pattern examples, finally, provide either an idiom containing the headword (e.g. "das wissen die Götter"—"the Gods will know" in the entry "Götter"—"Gods"—) or a schematic pattern illustrating its valency characteristics (e.g. "etwas erinnern"—"to remember sth"—, and "jemanden an etwas erinnern"—"to remind sb of sth" in the entry "erinnern").
- `<seg type="paraphrase">` marks paraphrases of illustrating examples (or parts thereof). Typically, these are provided for idiomatic or otherwise non-transparent uses of the headword. Thus, in the example "er hat in letzter Minute (ganz kurzfristig) abgesagt"—"he called off at the last minute (at very short notice)"—, the section in parentheses paraphrases an idiomatic use of the headword "Minute".[6]
- `<lbl>` marks various other types of information provided by the lexicographer. The preface of the WDG characterizes these stretches (included in a pair of slashes in the print version) vaguely as "grammatische und kommentierende Hinweise" ("grammatical and commenting hints"). They are clearly used as a kind of "miscellaneous" category serving a heterogeneous range of functions. One function is to replace or supplement a definition for headwords which are difficult to define through paraphrases, synonyms or antonyms (e.g. "/präzisiert eine Aussage/"—"specifies a statement"—, for the headword "vielmehr"). Another function is to subdivide lists of examples into literal, figurative, metaphoric or proverbial uses ("/bildl./," "/übertr./" and "/sprichw./," respectively). <lbl> elements also serve to supply semantic classes for nouns (e.g. "/Ländername/"—"country name"—) or for referents of adjectives (e.g. "/vom Menschen/"—"of humans"—) and to provide additional grammatical information not included in the <gramGrp> element of the entry (e.g. derivational information like abbreviations, diminutives etc. or case information for single words). Just like their function, the scopes of <lbl> elements vary. Some of them refer to an entire entry or sense, while others are valid only for a single example or even only for a single word in an example.
- `<usg>` marks usage information. As described in the dictionary preface, the WDG provides information about the headword's register (e.g. "dicht."—"poetic"— or "vulg."—"vulgar"—), its diachronic and diatopic classification (e.g. "hist."—"historic"—or "österr."—"Austrian"—) and its assignment to a special domain (e.g. "landw."—"agricultural"—or "Fußball"—"football"—). These specifications are frequently combined (e.g. "landw. österr.") and typically refer to the entire entry or sense, but they can also be used to characterize individual examples or parts thereof.
- `<w>` marks individual words with a special function or property. This comprises highlighted prepositions or stressed words in examples and citations of headwords in `<lbl>` elements (e.g. "/leitet in Anknüpfung an *eigentlich* eine Entgegnung ein/"—"initiates a riposte following (the word) *eigentlich*"—).
- `<ref>` marks links within the dictionary (e.g. references to related words as in "vgl. *Christus*" in the entry "Christ") and links from the dictionary to the list of sources. Using the latter links, it becomes possible to navigate from a single citation example to a list of all citations from the same work or the same author.

[6] Conceptually, there is no clear boundary between illustrating examples with a paraphrase element and pattern examples of idioms which are followed by a definition. Formally, however, the two can be distinguished unequivocally.

Other tags include <etym>, <gramGrp> and <form> for etymological, grammatical and form information about an entry, <pos> for part-of-speech labels, <cit> and <quote> for quoted examples and <bibl>, <author> and <title> for the corresponding bibliographic references.

4.2. *Transformation workflow*

The transformation is done in a semi-automatic process, matching patterns of typography-oriented mark-up and other structural indicators (mainly punctuation symbols such as colons, semicolons, parentheses etc.) and transforming them to corresponding content-oriented elements. For example, an italic passage at the beginning of a sense element, followed by a plain text passage starting with a colon and ending with a semicolon, can be reliably mapped to a definition, followed by a list of examples. However, owing to the size of the dictionary, the fine-graininess of the typographic mark-up (altogether there are roughly half a million typography-oriented tags in the source version) and the resulting combinatory possibilities, the number of different patterns which have to be mapped in this way is very large. This and the fact that a stepwise transformation process leads to complex interactions between source patterns which have not yet been treated and target structures which are not yet complete make the transformation a non-trivial task.

In order to ensure maximal accuracy, we have established the workflow depicted in figure 3: pattern candidates are identified by browsing the existing version of the dictionary and selecting recurring patterns with a clear content-oriented interpretation. Once a candidate has been chosen, it is described as an XPath expression and fed into the pattern transformer—an algorithm, implemented in Java and JDOM, consisting of three parts: the first part selects those instances in the source document which match a given pattern; the second part separates out possible exceptions;[7] the third part performs the actual transformation from source to target structure.[8] The whole entry or sense containing the target structure is then written into a separate document. This document can be viewed and checked with a browsing tool which displays the changed entries either as plain XML or as a (more easily readable) HTML visualization. Depending on the outcome of this check, the pattern is either abandoned (if it generates too many or too heterogeneous erroneous mappings) or modified (if it only generates systematic errors which can be caught either by modifying the pattern itself or by adding rules for handling exceptions), or the changed entries are written back into the original dictionary files, replacing the source entries (if there are no erroneous mappings). Several iterations of checking and modifying a pattern may be necessary before it is applied or abandoned.

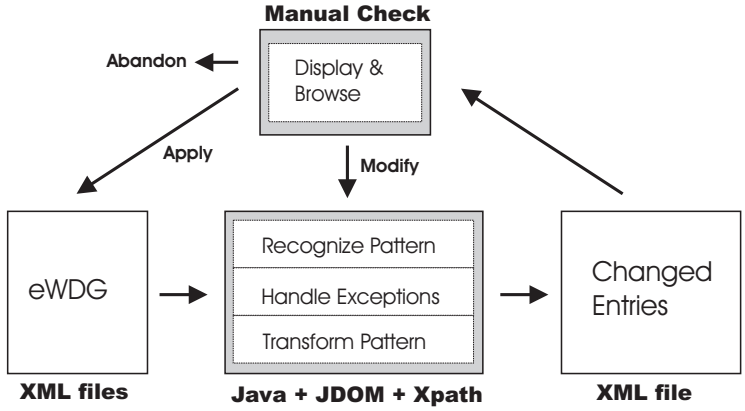

Figure 3: Transformation workflow

[7] This has proven to be more efficient than integrating the exceptions into the pattern itself.

[8] Each pattern transformer is implemented as an instantiation of an abstract class which takes care of most of the processing logic and contains methods for the most common transformation steps. In that way, writing a new pattern transformer is mostly a matter of specifying the parameters for the recognition and customizing the transformation methods.

The patterns vary greatly in complexity and in the number of instances they match in the dictionary files. Thus, about half of the dictionary entries can be completely transformed with no more than five simple patterns (these are all patterns of the form "usage information + definition + example(s)" or similar). It is only for longer entries that the complexity of the patterns increases and the instances they match decreases. Since more complex patterns are also more laborious to write and check, the number of completed entries grows more and more slowly as the transformation progresses (i.e. it takes about a tenth of the overall time to complete the first half of the entries, another tenth to complete the next quarter and so on). For the last 3,000 entries, it was more efficient to make the changes manually than to write and apply patterns that would transform no more than 10 or 20 entries.

The manual check was carried out on all changed entries if their number was smaller than 50. For greater numbers, only a sample was checked—typically 100 out of 1,000 changed entries—. Using this process, we are confident that the accuracy of the resulting mark-up is greater than 99%; i.e. in more than 99 out of 100 cases, we expect the new, content-based mark-up to fully reflect the actual content structure of the dictionary entry.

4.3. *Status and further steps*

The transformation of typography-oriented elements to the content-oriented elements listed in section 1.3 was completed in February 2008 after 8 months of work. Altogether, the entries and senses of the dictionary now contain around 90,000 definitions, 210,000 illustrating examples (25,000 of which contain a subordinate paraphrase element), 35,000 quoted examples, 10,000 pattern examples, 40,000 `<lbl>` elements, 135,000 `<ref>` elements and 45,000 `<usg>` elements. While we think that this is an adequate degree of detail for all the possible applications we are currently considering (see section 5), there are, of course, numerous possibilities of further refining the structure.

One additional refinement step could consist in adding a more hierarchical structure to the list of elements underneath a sense. For instance, many senses contain series of gradually more specific definitions, each with its own set of examples. Such elements could be grouped accordingly, as illustrated in figure 4.

```
<sense>
 <div>
   <!-- first definition: 'to collect money' -->
   <def>Geld kassieren</def>
   <eg>Beiträge a.</eg>
 </div>
 <div>
   <!-- second definition: 'to collect contributions, travel fare from all' -->
   <def>von allen den Beitrag, das Fahrgeld kassieren</def>
   <eg>die Mitglieder, Fahrgäste a.</eg>
   <eg>der hintere (Straßenbahn)wagen ist schon abkassiert</eg>
 </div>
</sense>
```

Figure 4: Grouping of elements in a sense of the headword "abkassieren" ("to cash up")

However, a potential difficulty in this refinement lies in determining the boundaries of a given group. Thus, an element initiating a certain group (e.g. the definition in the above example) may have in its scope all elements until the next initiating element of the same type, or it may hold only for the immediately following element. Sometimes, such boundaries are marked explicitly by a special structural indicator, but this is not always the case. Possibly, we encounter here a case where a structural feature of the dictionary is not (or not always) represented by an explicit indicator in the text. Our semi-automatic method would then meet its limits.

```
<sense>
 <!-- superordinate definition: 'collective housing, dwelling' -->
 <def>gemeinschaftliche Unterkunft, Wohnstätte</def> […]
 <sense>
   <!-- subordinate definition: 'for a certain group of people' -->
   <def>für einen bestimmten Personenkreis</def>
   <eg> das Kind wuchs in einem H. auf</eg> […]
 </sense>
 <sense>
   <!-- subordinate definition: 'for people seeking recreation, recovery' -->
   <def>für Erholungsuchende, Genesende</def>
   <eg> der Betrieb hat ein H. an der Ostsee</eg> […]
 </sense>
</sense>
```

Figure 5: Elliptic definitions in different senses of the keyword "Heim" ("home")

Another difficulty for a further refinement is of a more fundamental nature: since entries in the original WDG were intended to be read as a coherent whole, adding further (hierarchical) structure can lead to entities which are not fully interpretable in isolation. For instance, as figure 5 demonstrates, it is not uncommon for nested senses to use ellipses in which a subordinate definition supplements a superordinate definition with, e.g., a prepositional phrase.

It is hardly possible to detect such cases automatically, and it is questionable whether they can be transformed in such a way that subordinate elements become interpretable autonomously. However, especially in view of the various exploitation options sketched in the next section, this property of the dictionary has to be borne in mind, and any further refinements will have to make sure that divisions on lower levels respect and maintain the coherency at higher levels.

5. Exploiting the refined mark-up

The refinement described in the previous section opens various new or improved ways of visualizing and querying the dictionary's content. It also paves the way for a detailed linguistic and lexicographic analysis of the WDG.

5.1. *Visualization*

As regards visualization, the refinement of the mark-up is a necessary prerequisite for enhancing the readability of the dictionary. On the basis of the newly introduced content elements, various techniques can be applied to make the structure and content more easily accessible to the reader, for instance:

- Information of different types can be visually distinguished by appropriate formatting (e.g. definitions in bold, examples in italic) and/or by the use of appropriate layout elements (e.g. a new line for each example, indentation of embedded senses, a border around the definition of a subsense and its corresponding examples).
- Entries can be made more concise by showing or hiding (or expanding/condensing) certain types of information on demand (e.g. show only the first three examples for an overview, show more examples on user-demand).
- Entries or parts thereof can be rearranged to make salient information appear at more prominent places, e.g. phraseologisms containing the headword can be listed at the beginning of an entry.

As an example, compare the original entry "braten"—"to fry"—in the printed WDG and the eWDG.1 (figure 6) with an HTML visualization generated on the basis of the refined mark-up (figure 7). In the new visualization, general information about the entry is represented in a separate layout element (top left). The different pieces of information belonging to individual senses are structured by appropriate layout and formatting elements (middle and bottom left). On the right hand side, several abbreviated versions of the entry are provided, each one concentrating on a specific type of information: the first element presents a compact view of the

entry in which only the most salient element of each sense (a definition and an idiomatic expression, respectively) is included. The second element lists all pattern examples, the third all examples with paraphrases, and the fourth all quoted examples. Letting the mouse hover over these elements will display more information (e.g. number of definitions and examples, quoted passage) in a tooltip.

braten (er brät), briet, hat gebraten

1. *eine unfertige Speise im zerlassenen Fett in der erhitzten Pfanne mürbe, gar werden lassen*: eine Ente b.; den Fisch braun, knusprig b.; einen ganzen Ochsen am Spieß b.; sie brät sich /*Dat.*/ schnell ein Schnitzel; etw. am kleinen Feuer, über schwacher Flamme, in Butter b.; jetzt können wir Äpfel braten in unserem Ofen SEGHERS 4,214 (*Siebtes Kreuz*); gebratene Zwiebel; /*übertr.*/ salopp jmdm. eine Extrawurst b. (*jmdm. eine unverdiente Bevorzugung zuteil werden lassen*); die gebratenen Tauben fliegen niemandem ins Maul (*jeder muß sich anstrengen*); da brat mir einer 'nen Storch /*Ausruf der Verwunderung, Entrüstung/ das ist seltsam*!

2. etw. brät *etw. wird im zerlassenen Fett in der erhitzten Pfanne mürbe, gar*: die Ente brät schon, wird wohl zwei Stunden b. müssen, brät langsam; die Kartoffeln b. im Tiegel; /*übertr.*/ salopp er muß in der Hölle b.

braten (er brät), briet, hat gebraten

1. *eine unfertige Speise im zerlassenen Fett in der erhitzten Pfanne mürbe, gar werden lassen*: eine Ente b.; den Fisch braun, knusprig b.; einen ganzen Ochsen am Spieß b.; sie brät sich /*Dat.*/ schnell ein Schnitzel; etw. am kleinen Feuer, über schwacher Flamme, in Butter b.; jetzt können wir Äpfel braten in unserem Ofen Seghers 4,214 (*Siebtes Kreuz*); gebratene Zwiebel; /*übertr.*/ salopp jmdm. eine Extrawurst b. (*jmdm. eine unverdiente Bevorzugung zuteil werden lassen*); die gebratenen Tauben fliegen niemandem ins Maul (*jeder muß sich anstrengen*); da brat mir einer 'nen Storch /*Ausruf der Verwunderung, Entrüstung/ das ist seltsam*!

2. etw. brät *etw. wird im zerlassenen Fett in der erhitzten Pfanne mürbe, gar*: die Ente brät schon, wird wohl zwei Stunden b. müssen, brät langsam; die Kartoffeln b. im Tiegel; /*übertr.*/ salopp er muß in der Hölle b.

Figure 6: Presentation of the entry "braten" ("to fry") in the printed WDG (left) and on the basis of the eWDG.1 mark-up (right)

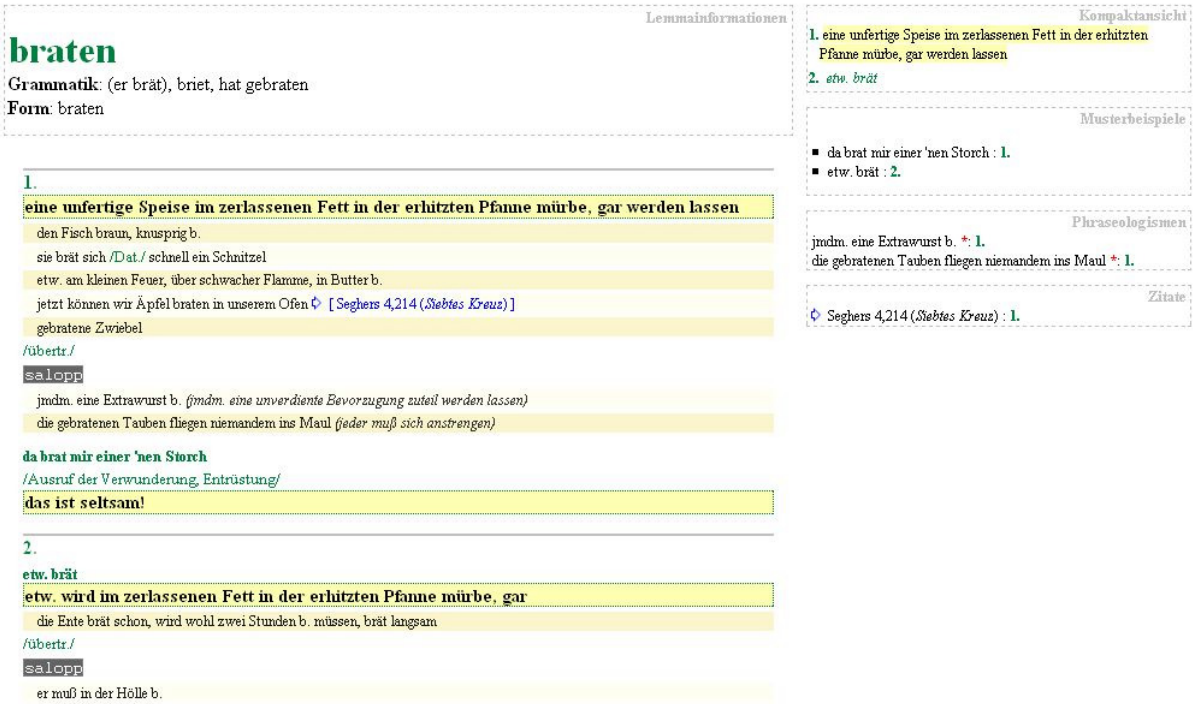

Figure 7: Web presentation of the same entry on the basis of the eWDG.2 mark-up (abbreviated to save space)

With similar techniques, the visualization of entries can be adapted and optimized for a specific viewer, e.g. a web browser vs. a mobile device, and/or for a specific user scenario, e.g. a language learner vs. a professional translator.

5.2. *Querying and analysis*

Similarly, the refined structure can be exploited for dictionary querying. In version 1 of the eWDG, querying could be done alternatively as a simple headword lookup or as full text search of entries, the first option leaving much information in the dictionary unused, the second potentially leading to very heterogeneous and thus unwieldy search results. On the basis of the new content-based mark-up, text searches can now be restricted to certain types of information. As an example, consider a text search for the word "Meinung" ("opinion"). When restricted to `<seg type="paraphrase">` elements (i.e. paraphrases of illustrating examples), the result will consist mainly of (typically multi-word) expressions paraphrasing "Meinung (sagen)"—"(to state an) opinion"—, whereas a query on `<eg>` elements (i.e. illustrating examples) is most useful for identifying elements collocating with "Meinung". A query on `<def>` elements, finally, yields a more heterogeneous result, containing synonyms, hyponyms and words for otherwise semantically related concepts. Figure 7 lists some example results for each type of query.

<seg type="paraphrase"> elements (50 more hits)	Headword
dann hat er ordentlich *ausgepackt* (deutlich seine *Meinung* gesagt)	auspacken
den habe ich richtig *bedient* (dem habe ich die *Meinung* gesagt)	bedienen
jmdm. *Bescheid* sagen (jmdm. die *Meinung* sagen)	Bescheid
die *Katze* aus dem Sack lassen (seine wahre *Meinung* zeigen)	Katze
jmdm. (gehörig) den *Marsch* blasen (jmdm. unmißverständlich die *Meinung* sagen)	Marsch
den *Mantel* nach dem Winde drehen (seine *Meinung* je nach Vorteil ändern)	Mantel
<eg> elements (76 more hits)	
jmdm. seine *Meinung aufoktroyieren*	aufoktroyieren
seine *Meinung* […] (freimütig, unumwunden, unverhohlen) zu einer Frage, über jmdn. *Äußern*	Äußern
ihre Ansichten, *Meinungen divergieren*	Divergieren
<def> elements (128 more hits)	
Ansicht, Meinung	Auffassung
jmd., der nie Widerstand zu leisten oder seine *Meinung* offen zu sagen wagt, Duckmäuser	Leisetreter
eine gegensätzliche *Meinung* vertreten, jmdm. widersprechen, sich jmdm. widersetzen	Opponieren
jmd., der ständig eigensinnig eine andere *Meinung* vertritt oder anders handelt	Querkopf
durch Erfahrung, Prüfung gefestigte *Meinung* von dem, was wahr, richtig ist, […]	Überzeugung

Figure 8: Text queries for "Meinung" restricted to different element types

Last but not least, the refined mark-up can also be exploited for linguistic and lexicographic studies of the WDG. Just as a dictionary user is given new options for querying and viewing the dictionary's content, linguists and lexicographers can use the more detailed distinction of content elements to carry out analyses with and about the dictionary. Thus, the already existing calculation of synonyms, hyponyms and hypernyms can be refined and improved, because reliable information about which parts of the entry text are definitions and which are not is now available. Similarly, the abundance of illustrating examples, which are now reliably recognizable as such, offers various options for the analysis of collocations and valency patterns. As the project progresses, we plan to integrate the results of such analyses into the digital lexical system and make them available to the dictionary user.

6. Conclusion and outlook

In this paper, we have shown how we transformed typography-based mark-up of dictionary articles into content-based mark-up, and we have explained why and how this transformed mark-up is put to use in the context of a lexical information system. We think that this is a task which will arise in many projects that aim to transfer printed lexical resources to a digital environment. While the details of the task may differ from case to case, some more general observations can be made: first, the task is not necessarily one of adding information to the resource, but rather of making implicit information, contained in formatting and structural

indicators, explicit. Second, the regularities on which this process is based, though simple in individual instances, become complex when applied to the dictionary as a whole. This makes it necessary to mix automatic recognition and transformation methods with manual checking and correction procedures. The task thus becomes more costly but, we think, remains manageable with a justifiable amount of time and effort. Third, the refined structure can be used to improve visualization and query of the dictionary resource. If, as is the case in our project, the dictionary is a part of a larger lexical information system, the new structural elements will also be used as the basis for interlinking and extending different resources in that system.

Work in the near future will be concerned, on the one hand, with additional refinements of the article structure. Most importantly, we are currently exploring methods of making inter-article links (e.g. from a derived adjective to its base noun) and links from the dictionary to external sources (e.g. from a quoted example to a list of bibliographical sources), systematically accessible. On the other hand, we are preparing a new online version of the DWDS in which the new features described in this paper, alongside other improvements and extensions of the dictionary and the corpus components of the lexical information system, will be made available to the user.

References

Alshawi, H.; Boguraev, B.; Carter, D. (1989). "Placing the dictionary on-line". In Boguraev, B.; Briscoe, T. (eds.) *Computational Lexicography for Natural Language Processing*. London, New York: Longman. 41-64.

Fellbaum, C. (2004). "Idiome in einem Digitalen Lexikalischen System". *Zeitschrift für Literaturwissenschaft und Linguistik* 136.

Fellbaum, C. (2007) (ed.) *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum Press.

Geyken, A. (2007). "The DWDS corpus: A reference corpus for the German language of the 20th century". In Fellbaum, C. (ed.) *Collocations and Idioms.* London: Continuum Press. 23-40.

Geyken, A. (2005). "Das Wortinformationssystem des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS)". *BBAW Circular* 32. Berlin: BBAW.

Geyken, A.; Hanneforth, T. (2006). "TAGH: A complete morphology for German based on weighted finite state automata". In *Proceedings of FSMNLP 2005, Lecture Notes on Artificial Intelligence*.

Geyken, A.; Ludwig, R. (2003). "Halbautomatische Extraktion einer Hyperonymiehierarchie aus dem Wörterbuch der deutschen Gegenwartssprache". *TaCoS 2003, Gießen* 13.-15.6.

Hauser, R.; Storrer, A. (1994). "Dictionary Entry Parsing Using the LexParse System". In *Lexikographica 9/1993*. Tübingen. 174-219.

Klein, W. (2004a). "Vom Wörterbuch zum Digitalen Lexikalischen System". *Zeitschrift für Literaturwissenschaft und Linguistik* 136.

Klein, W. (2004b). "Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts". In Scharnhorst, J. (ed.) *Sprachkultur und Lexikographie*. Frankfurt am Main: Peter Lang. 281-309.

Klein, W.; Geyken, A. (2000). "Projekt, Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts". *Jahrbuch der BBAW 1999*. Berlin: Akademie-Verlag. 277-289.

Malige-Klappenbach, H. (1986). *Das Wörterbuch der deutschen Gegenwartssprache: Bericht, Dokumentation und Diskussion*. Tübingen: Niemeyer.

Neff, M. S.; Boguraev, B. K. (1989). "Dictionaries, dictionary grammars, and dictionary entry parsing". In *Proceedings of the 27th ACL, Vancouver*. 91-101.

Schiller, A.; Teufel, S.; Stöckert, C. (1999). "Guidelines für das Tagging deutscher Textcorpora mit STTS". *Research report*. Stuttgart, Túbingen: Universität Stuttgart und Universität Tübingen.

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. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*. Tübingen: Niemeyer. 88-104.

Wiegand, H. E. (1990). "Die deutsche Lexikographie der Gegenwart". In Hausmann, F. J. et al. (eds.) *Wörterbücher. Ein internationales Handbuch zur Lexikographie, 2. Teilband*. Berlin, New York: de Gruyter. 2100-2246.

Dictionaries

[DWB]. *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm*. 16 Bde. [in 32 Teilbänden]. Leipzig: S. Hirzel. 1854-1960.

[WDG]. Klappenbach, R.; Steinitz, W. (eds.) (1964-1977). *Wörterbuch der deutschen Gegenwartssprache (WDG)*. 6 Bände. Berlin: Akademie-Verlag.