Integrating Lexical Resources Through an Aligned Lemma List

Axel Herold, Lothar Lemnitzer, and Alexander Geyken

Abstract This paper presents the modelling of a common meta-index for large modern and historical lexical resources of the DWDS project. Due to the different lexicographical principles and traditions employed for these resources as well as the different historical periods covered, such a meta-index cannot be modelled as a simple list of 1 : 1-correspondences between entries across different dictionaries. In order to model the occurring phenomena such as graphematic headword variance, homography, semantic change and differences in the semantic entry structure a more complex typed link structure is required.

1 Project Background

We report on one facet of a long-term project which aims at the integration of several lexical and textual resources in order to document the German language and its use at several stages. The integration affects the dictionaries which are procured at the project *Digitales Wörterbuch der deutschen Sprache* (DWDS),[1] a project of the *Zentrum Sprache* at the Berlin-Brandenburg Academy of Science and the Humanities (Klein and Geyken, 2010). The synchronization of various dictionaries via a common index enables the simultaneous querying and display of information which can be considered to be information about the same lexical entry. The possibility to formulate queries across its data sources – dictionaries, corpora, statistical tools – is the key feature of the lexical information system DWDS (see Klein, 2004 for a detailed discussion of the lexical information system).

Axel Herold · Lothar Lemnitzer · Alexander Geyken
Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, D-10117 Berlin
e-mail: {herold,lemnitzer,geyken}@bbaw.de

[1] http://www.dwds.de/

The DWDS comprises a broad range of lexical resources:

eWDG2 is a faithful digitized version of the *Wörterbuch der deutschen Gegenwartssprache* (WDG, 1962-1977) a six-volume printed dictionary compiled at the *Akademie der Wissenschaften der DDR*. It contains 120 000 entries and is represented in TEI-P5 compliant markup (Herold, 2011).

DWDSWB is a new and continuously extended edition of the WDG which – in addition to its WDG heritage – contains more contemporary German vocabulary, i.e. lexical entries which have come into use after 1977. New entries for about 25 000 headwords will be compiled by the DWDS project team during the next six years.

EtymWB is a faithful digitized version of the three-volume *Etymologisches Wörterbuch des Deutschen* (Pfeifer, 1989), a printed dictionary compiled at the *Akademie der Wissenschaften der DDR* in the 1980s. It contains entries for 8 000 morphologically simple main headwords and 14 000 derived minor headwords.

[1]DWB is a faithful digitized version of the first edition of the *Deutsches Wörterbuch* (DWB, 1854-1961). This 33-volume printed dictionary contains about 300 000 main headwords and a yet unknown number of related minor headwords. For a documentation of the editing process of [1]DWB see Dückert (1987); Schmidt (2004) provides a helpful introduction to the (proper) use of this dictionary.

Besides these lexical resources, the DWDS provides:

- a time and genre balanced corpus of the German language of the 20th/21st century which comprises 100 million tokens for the period 1900–2000 and roughly ten million tokens for the first decade of the 21st century (Geyken, 2007);
- an extended, opportunistic corpus mostly compiled from electronic newspaper texts with a total of 2.7 billion tokens of which approximately one billion tokens are publicly accessible;
- statistical results of corpus analysis, e.g. word distribution patterns and a word-sketch like application (*Wortprofil*, see Geyken et al, 2009).

The corpus resources are searchable with DDC, a powerful linguistic search engine (Sokirko, 2003). With the exception of [1]DWB, which will be published towards the end of 2012, all data are searchable on-line at http://www.dwds.de/. Each resource is displayed in its own panel, that accounts for the specific presentational needs of the resource.

2 Linking Across Dictionaries

The DWDS project aims at providing explicit links across all its lexical resources on the level of entries. Links between lexical resources and corpora are automatically established on the basis of lemmas and are subject to changes when the linguistic (pre)processing of the corpora is improved. In this paper we focus on the construc-

tion and maintenance of explicit linkage, i.e. on constructing and representing a list of equivalent lemmas.

2.1 The Building Blocks: Entries, Lemmas, Headwords

For the lexical resources the main access path is provided on the level of individual lexical entries. Each lexical entry E describes one or more lemmas L. Lemmas in turn are represented by one or more headwords H.

$$E := \{L_1, \ldots, L_n\} \; ; \; L := \{H_1, \ldots, H_m\} \tag{1}$$

Throughout this paper, headwords are noted exactly as they appear in the printed source, e.g. $^{1\text{DWB}}$SONNTAG, $^{\text{WDG}}$Sonntag ‘Sunday’.

WDG and DWDSWB constrain their entries to contain a single lemma. In EtymWB entries etymologically related lemmas are grouped together and are typically described discursively. This discursive structure often prohibits splitting entries into smaller self-contained parts describing one lemma only. Due to the long editing period of ^{1}DWB there are many inconsistencies with regard to the structure of lexical entries. The majority of entries is self-contained and understandable without context regardless of their morphological complexity. However, there are instances of entry embedding, i.e. entries being located inside other entries where the exact location of the sub-entry cannot be ignored because it associates the sub-lemma with a distinct sense of the main lemma (e.g. see $^{1\text{DWB}}$SONNTAG ‘sunday’ II 2 e). We follow a liberal approach to identifying entries in ^{1}DWB: They must contain at least one headword (not necessarily entry initial), they have to be self-contained and we try to minimize the size of entries, i.e. we try to maximize the number of entries.

2.2 Entry Equivalence

The primary goal of linking the lexical resources explicitly is to reach high accuracy in the matching of lexical entries of the involved resources. The user should see in the dictionary panels the information found in the selected dictionaries for exactly the same lemma. We therefore opted for a manually controlled automatic alignment process. The result is an aligned lemma list which contains the lemma equivalence relations across all lexical resources.

As long as only the ‘modern’ contemporary dictionaries DWDSWB, WDG and EtymWB are considered, a binary equivalence function is sufficient to express whether two lemmas are equivalent or not. Lemma and headword selection target the same user group and the dictionaries were created around the mid/end of the 20th century. Based on this notion of equivalence the EtymWB can easily be ex-

ploited as an etymological supplement to the DWDSWB and WDG. Most of the alignment between the three dictionaries could be achieved automatically based on headwords because the headword selection scheme is nearly identical.

There were some challenges, though, that required manual correction to differing degrees:

Homography/homonymy: Homonymy – or homography, its lexicographical incarnation – is a long-debated topic in lexicology and lexicography. Up to now, researchers only agree that there are various reasons to assign homonymy to a headword, the relevant reasons among them are (see Behrens, 2002 for a more complete overview):

1. the lexical entries which are represented by a common headword have formal differences, e.g. in gender, inflection or conjugation (formal criterion);
2. the lexical form has meanings which are completely unrelated, even from a diachronic point of view (semantic or etymological criterion).

These criteria are orthogonal and each of them is based on concepts which give room to interpretation. Consequently they lead to some variance in the lexicographic practice even in the case where one criterion is followed explicitly. For an overview of the lexicographical practice in various synchronic German dictionaries see Kempcke (2001). The WDG follows the formal criterion and applies it rather modestly (WDG, 1962-1977, p. 20 f.; Kempcke, 2001, p. 63 f.). While in the EtymWB's preface no mention is made of any criterion, it might go without saying for an etymological dictionary that the etymological criterion is applied and the same holds for [1]DWB. To give just one example: WDG establishes two lexical entries for [WDG]See, i.e. *der See* (masculine) 'lake', and *die See* (feminine) 'sea'. As there is no diachronic or etymological reason to tear these meanings apart, hence there is only one lexical entry in EtymWB and in [1]DWB respectively.

Since there is no single explicit criterion for ordering lexical entries with identical headwords, there is variance in the order of homograph entries between the dictionaries.

Incompatibility: By this term we mean that one dictionary supplies more homographs (entries) for a headword than the other. For example, EtymWB (and again, [1]DWB) presents a reading of [EtymWB]Art, roughly meaning 'ploughed land', which is not used in contemporary German and therefore missing in the WDG and DWDSWB. On the other hand, the WDG lists two homographs for [WDG]ober 'above' (as does [1]DWB), one being a variant which is only used in Austria, while EtymWB has only one entry with that headword.

Headword variance: According to some variance in selecting headwords that represent the lemma, we can find differing headwords for obviously identical lemmas across dictionaries. Often the cause for this is a regular choice among canonical forms, e.g. for de-adjectival nouns as in *(ein) Angesteller* vs. *(der) Angestellte* 'clerk' (indefinite vs. definite determiner). Another reason for headword variance lies in the strong tendency for the use of the plural form of a noun while the singular form is not completely ruled out and might even have been predominant in

earlier language stages. This is the case with e.g. *Aliment* (singular) vs. *Alimente* (plural) 'alimony'.

Most of these issues presented obstacles to a straight-forward automatic alignment of entries which would have been based on the form of the headword alone.

We are currently working on the integration of [1]DWB, which poses some extra challenges on top of those which have been described above:

Homography/homonymy: Again, we face the consequences of different theories of homonymy across dictionaries leading to different numbers of homographs for a given lemma. Additionally, homographs in [1]DWB are not consistently marked and in some cases not marked at all (e.g. [1DWB]MAST with two entries for a male noun and two for a female noun). Currently there are more than 10 000 headwords known to appear multiple times in [1]DWB. As still more headwords are identified these numbers are even expected to increase.

Semantic change: The first volume of [1]DWB was issued in 1854 – more than 100 years before the first volume of WDG appeared in 1962. Consequently there is a considerable amount of lexical entries where the description of formally identical lemmas clearly shows radical semantic changes leeding to incongruent or even incompatible semantic descriptions. Let us consider a selection of lemmas that are recorded as non-homographic in [1]DWB and WDG:

- [1DWB]BARBAR vs. [WDG]Barbar – while [1]DWB describes the historical Greek meaning of the lemma only ('foreign people, foreign human, non-Greek'), we find three distinct sense descriptions in the WDG of which only the last one (explicitly marked as historical) corresponds with the sense description given in [1]DWB. The other two modern senses ('cruel human' and 'ignorant people, philistine') are derived from metaphorical uses of *Barbar* and are predominant in present-day German.
- [1DWB]GEBILDET vs. [WDG]gebildet – here, [1]DWB lists three senses for the lemma, namely 'illustrated', 'shaped, made of' and 'educated, intellectual' of which only sense number three is directly attested in the WDG. Following the pointer to the derivational base *bilden* which is also given in the WDG, the second sense can be deduced by the reader even though it is not stated explicitly. However, the first sense attested in [1]DWB remains exclusive to this dictionary.
- An example for a complete meaning shift is [1DWB]TRILLION vs. [WDG]Trillion where the former describes it as *tausend billionen* (10^{15}) and the latter describes it as 10^{18} Unfortunately there is no semantic paraphrase for *billion* in [1]DWB and some other powers of ten like *billiarde* do not appear as headwords at all. This phenomenon of incompatible meanings is also frequently found in the semantic description of man-made artifacts like [1DWB]HOLZSTOCK 'chopping block' and 'printing block' vs. [WDG]Holzstock 'wooden stick'.

Under- or even unregulated orthography: Special forms of headword variance are due to the hazards of an under-regulated, if not unregulated orthography at least in the first half of the time in which [1]DWB was compiled.

In cases where orthographical changes lead to a headword appearing under another initial letter again, "continuation entries" were inserted effectively resulting in two entries describing exactly the same lemma. The continuation typically consists of complementary information. Consider for example $^{1\text{DWB}}$CAPELLE 'chapel'. This entry appeared in an 1855 partial issue, before (in an 1864 partial issue) $^{1\text{DWB}}$KAPELLE 'chapel' was introduced pointing to the older entry and extending it with a five-fold explicit sense description and many more usage examples.

In ^{1}DWB, there are even occurrences of headwords that are judged as orthographic "errors" by the editors which didn't stop them from writing an article about that lemma, e.g. $^{1\text{DWB}}$CAPELLE *fehlerhaft für cupelle, cupella* 'wrongly (written) for cupelle, cupella'.

The rather late fixing of orthographical norms – the first official norm for written German was decided in 1902 –, as well as various changes of this norm since then, cause differences in the form of headwords which are hard to capture. One typical pattern is the letter sequence *th* that was replaced by *t* in many (but not all) words, e.g. *Thal* $\longrightarrow$ *Tal* 'valley'. Another frequent pattern is the variance between *c* and *k*. There are regular patterns of orthographic change which we will try to capture with an orthographic normalizer. We use CAB (cascaded analysis broker) for this purpose, a rule based transducer which originally maps historical to contemporary spelling (Jurish, 2010). The technology, however, allows us to use the program the other way round and to produce possible historical spelling variants of contemporary headwords. The overgeneration of forms can be controlled by matching all output forms with the ^{1}DWB headwords list.

However, there are still many so-called "false friends" which are an obstacle to any automatic alignment. Let us illustrate this point with an example: Naïve alignment would run into problems with headwords like $^{\text{DWDSWB}}$Turm. There is a corresponding headword in ^{1}DWB, but this one represents another lemma (which is not in use in contemporary German). The correct correspondence is $^{1\text{DWB}}$THURM.

Idiosyncratic canonicalization of headwords: For reasons which are hardly understandable nowadays, the lexicographers chose to use canonical forms as headwords which look quite idiosyncratic to the contemporary user and lexicographer. First, all headwords in ^{1}DWB appear completely in capital letters, while in DWDSWB as in most other contemporary German dictionaries nouns start with a capital letter while words of other parts of speech do not, e.g. $^{\text{DWDSWB}}$Weg vs. $^{1\text{DWB}}$WEG 'way' and $^{\text{DWDSWB}}$weg vs. $^{1\text{DWB}}$WEG 'gone'. For a correct mapping of these entries, the part of speech information given in both dictionaries can be used, e.g. *masculine noun* vs. *adverb* for $^{1\text{DWB}}$WEG. To complicate matters, though, part of speech information is not available for all entries in ^{1}DWB.

The letter *ß* (eszett or sz-ligature) is specific to the German writing system. It is represented literally as *SZ* in ^{1}DWB headwords, e.g. $^{\text{DWDSWB}}$Spaß vs. $^{1\text{DWB}}$SPASZ 'fun, joke'. This is effectively a lossy transformation and cannot be simply reversed without morphological analysis because the letter sequence

sz is of course completely legal in German, e.g. [DWDSWB]Lebenszeichen ‘vital sign’.

Historical and dialectal lemmas: Due to the dictionary’s decidedly diachronic view there is a considerable amount of lemmas that were in use in historic times but are not attested in DWDSWB or EtymWB because they fell out of general use. Typically those lemmas cannot be found in present day corpora except, perhaps, as common names, e.g. [1DWB]SESTER, a historical measure of capacity.
Another group of lemmas that are not found in today’s general dictionaries are only used in certain dialects, many of which are not explicitly marked as such in [1]DWB, e.g. Lower German [1DWB]PADDEN ‘making small steps (like a toad)’.

Although formal variance in headwords representing the same lemma is an obstacle for automatic matching of entries across dictionaries this phenomenon can be accounted for by acknowledging different lemmatisation strategies and allowing for different orthographies. However, due to different accounts to homography and more so incongruent or even incompatible semantic descriptions an aligned lemma list cannot be modelled as a simple list of $1:1$ correspondences between lexical entries.

3 Evaluation

DWDSWB, WDG, and EtymWB are completely aligned and the alignment is actively exploited on the project’s website to achieve a synchronized display of equivalent lexical entries. This makes it possible to use EtymWB as an etymological extension to the synchronous view of the present-day dictionaries.

First alignment tests between DWDSWB and [1]DWB on a random sample (941 entries) of approximately 45 000 non-homographic entries appearing in both dictionaries show strong semantic equivalence for about 67 % (632) of those entries. Another 3 % (27) are clearly not semanticly related, i.e. their semantic descriptions are incompatible. For about 8 % (79) we found partial semantic overlap. These pairings of entries are instances of incongruent semantic descriptions. The remaining entries 22 % (203) cannot be evaluated based on the entries alone, typically because DWDSWB lists a considerable amount of lemmas only with their headword(s) and still lacks semantic descriptions for them. During the course of the DWDS project these omissions will be amended.

While the figures for headword identical entries already look promising there is still much space for improving the alignment strategy considering the total number of entries and headwords in both dictionaries. To achieve a wider coverage on the non-homographic entries we will use CAB as a filter to derive additional mappings between contemporary and historical headwords. These suggestions have of course to be accepted or dismissed by a skilled lexicographer. Our aim is still to cover as many entries of these resources as possible.

4 Reusable Representation

As a result of our alignment efforts we want to provide to the community a combined lemma list of our lexical resources which will encompass synchronic as well as diachronic dictionaries and which will enable other lexicographic data centres to align their resources with this list. There are different technical standards to represent an aligned lemma list in a reusable way:

RDF (Resource description framework): RDF is very generic in allowing to express relations among arbitrary entities. Currently we are working towards a taxonomy of entry relations based on the phenomena sketched in Sect. 2 that will allow to express the relations among the dictionaries in RDF statements.

LMF (Lexical markup framework): LMF is an emerging standard directed specifically towards the representation of lexical resources. There are two basic modelling options with regard to LMF:

- transferring DWDSWB into LMF and providing links to the other dictionaries (e.g. via LMF's multilingual notations extension) which in turn would not have to be modelled in LMF;
- considering the aligned lemma list as a resource in its own right and model it in LMF, again providing links to the dictionaries.

The second option allows for easier maintenance of headword variants and can be directly exploited for queries across the represented dictionaries. It allows linkage among arbitrary dictionaries because it does not require a 'central' dictionary to contain a semantically equivalent entry. Most importantly, it provides a clear distinction between the underlying lexical resources and the relations among them similar to the RDF representation.

The aligned lemma list will be made available in RDF and LMF.

5 Future Work

We clearly need a solid operationalisation for the concept of semantic equivalence that allows for robust classification. Given the huge amount of manual effort needed to complete the alignment between DWDSWB and [1]DWB on the level of lexical entries it seems unfeasible to achieve a mapping for individual senses.

The lexical entries of DWDSWB, WDG, EtymWB and [1]DWB will become publicly exposed through the project CLARIN (Common Language Resources and Technology Infrastructure) where persistent identifiers will be employed to provide a stable reference system across different versions of the dictionaries.

In the near future we are also going to extend the aligned lemma list with headwords from GermaNet (Kunze and Lemnitzer, 2010). Since GermaNet is a resource which orders headwords according to individual senses, we expect further challenges to the alignment.

Acknowledgements This work was supported by the long term project DWDS of the Berlin-Brandenburg Academy of Science and the Humanities and CLARIN-D (Common Language Resources and Technology Infrastructure, http://de.clarin.eu/), funded by the German Federal Ministry for Education and Research.

References

Behrens L (2002) Structuring of word meaning II: Aspects of polysemy. In: Cruse DA, Hundsnurscher F, Job M, Lutzeier PR (eds) Lexikologie – Lexicology. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen, vol 1, de Gruyter, Berlin, pp 319–337

DWB (1854-1961) Deutsches Wörterbuch. Hirzel, Leipzig

Dückert J (ed) (1987) Das Grimmsche Wörterbuch. Untersuchungen zur lexikographischen Methodologie. Hirzel, Leipzig

Geyken A (2007) The DWDS corpus: A reference corpus for the German language of the twentieth century. In: Fellbaum C (ed) Idioms and collocations: Corpus-based linguistic and lexicographic studies, Research in corpus and discourse, Continuum, London, pp 23–40

Geyken A, Didakowski J, Siebert A (2009) Generation of word profiles for large German corpora. In: Kawaguchi Y, Minegishi M, Durand J (eds) Corpus analysis and variation in linguistics, Studies in Linguistics, vol 1, John Benjamins, pp 141–157

Herold A (2011) Retrodigitalisierung und Modellierung des Wörterbuchs der deutschen Gegenwartssprache. In: Krafft A, Spiegel C (eds) Sprachliche Förderung und Weiterbildung – transdisziplinär, no. 51 in Forum angewandte Linguistik, Peter Lang, Frankfurt (M.), Berlin

Jurish B (2010) More than words. Using token context to improve canonicalization of historical German. JLCL 25(1):23–40

Kempcke G (2001) Polysemie oder Homonymie? Zur Praxis der Bedeutungsgliederung in den Wörterbuchartikeln synchronischer einsprachiger Wörterbücher der Deutschen Sprache. Lexicographica 17:61–68

Klein W (2004) Vom Wörterbuch zum Digitalen Lexikalischen System. Zeitschrift für Literaturwissenschaft und Linguistik 136:10–55

Klein W, Geyken A (2010) Das digitale Wörterbuch der deutschen Sprache (DWDS). Lexicographica 26:79–93

Kunze C, Lemnitzer L (2010) Lexical-semantic and conceptual relations in GermaNet. In: Storjohann P (ed) Lexical-semantic relations: Theoretical and practical perspectives, no. 28 in Lingvisticæ Investigationes Supplementa, John Benjamins, Amsterdam, pp 163–183

Pfeifer W (1989) Etymologisches Wörterbuch des Deutschen. Akademie-Verlag, Berlin

Schmidt H (2004) Das Deutsche Wörterbuch. Gebrauchsanweisung. In: Bartz HW, Burch T, Christmann R, Gärtner K, Hildenbrandt V, Schares T, Wegge K (eds)

Deutsches Wörterbuch. Elektronische Ausgabe der Erstbearbeitung von Jacob Grimm und Wilhelm Grimm, Zweitausendeins, Frankfurt (M.), pp 25–64

Sokirko A (2003) DDC – a search engine for linguistically annotated corpora. In: Proceedings of Dialog 2003, Protvino (Russia)

WDG (1962-1977) Wörterbuch der deutschen Gegenwartssprache. Akademie-Verlag, Berlin