Digital and traditional resources for the second edition of the *Deutsches Wörterbuch*

Elke Gehweiler & Christiane Unger

Keywords: *diachronic lexicography, Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, digital resources, historical corpus, quotation archive*.

Abstract

The paper gives a short overview of the selection of sources for the first and second editions of the Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, from which the quotations for the dictionary entries are drawn. We will introduce the 2DWB quotation archive, which forms the basis of the lexicographical work on the second edition of the Deutsches Wörterbuch (2DWB) and which today is complemented by digital resources. We will assess a number of freely available digital collections of text according to their suitability for diachronic lexicography. We will look at size, selection of texts, verifiability of search results, quality of full texts and scans, presentation of search results and search functions. It will turn out that none of the resources can (yet) substitute 2DWB archive. We will further suggest that from the point of view of diachronic lexicography in some areas the examples from the "intelligent" quotation archive are superior to automatically retrieved examples from digital corpora.

1. Introduction

The *Deutsches Wörterbuch* is the big historical dictionary of German. It was founded in 1854 by Jacob Grimm and Wilhelm Grimm and was only finished more than 100 years later, in 1961, when the last of altogether 32 volumes appeared. The first edition (^{1}DWB) contains around 330.000 headwords, which makes it by far the most comprehensive dictionary of German. The brothers Grimm themselves only managed to get as far as the letter F: Wilhelm died in 1859, Jacob in 1863 while working on *Frucht* 'fruit'. A digital version of ^{1}DWB is available online (http://woerterbuchnetz.de/DWB).

Even before the last volume of the first edition was completed, it was clear that a revision of at least the most outdated volumes would be necessary and already in 1961 work on a new edition for the letters A-F started. The second edition (^{2}DWB) is currently being compiled at the Academies of Sciences and Humanities in Berlin and Göttingen and is about to be completed. Its digitization is currently being prepared.

The present paper will first describe the sources for the first and second editions, from which the quotations for the dictionary entries were drawn. The digital resources of today have to live up to the same standards as the older sources, but they have to fulfill a number of additional criteria if they are supposed to make historical lexicography easier and faster. We will assess a number of freely available collections of German historical text with regard to their suitability for historical lexicographic research. We will further address the differences in quality between automatically retrieved and manually excerpted examples.

2. The quotation archive

In the preface of the first edition Jacob Grimm established the principles that would henceforth guide the work of many historical lexicographers (cf. Kirkness 2012). These principles concerned the aim and scope of the dictionary, the presentation of information in dictionary entries, the criteria for inclusion in the dictionary and the selection of sources for the dictionary. Regarding the last point Jacob Grimm claimed that every statement should be substantiated by quotations. He writes: “Wörter verlangen beispiele, die beispiele gewähr, ohne welche ihre beste kraft verloren gienge. Wie könnten stellen (loci) heiszen, deren stelle ungenannt bliebe? Der name ihres urhebers reicht nicht aus, sie müssen aufgeschlagen werden können; ... unbelegte citate sind .. unbeglaubigte, unbeeidete zeugen.“ (^{1}DWB 1, XXXVI)

The Grimms took care to achieve a historical and regional balance of sources (cf. ^{1}DWB 1, IV-V), with a focus on linguistically influential authors like Luther, Hans Sachs, Geiler von Keisersberg and important writers like Opitz, Gellert, Lessing or Goethe and Schiller. They also included non-fiction, for example scientific texts, texts representing special languages or newspapers (cf. Dückert 1987: 34 ff.). These sources where excerpted in large reading programmes. The bibliographies for the second and third volumes of ^{1}DWB show that the resources for the dictionary were reassessed and expanded regularly; the bibliography for the entire first edition, finished in 1971, lists more than 25.000 titles.

The quotation archive of the new edition is based on the bibliography of its predecessor. It contains around 6.5 million quotation slips, drawn from more than 10.000 sources covering all periods of German, 6.000 of which were excerpted systematically. Although there are no more reading programmes today, the archive is still updated, though not systematically. It can be accessed through an electronic lemma list, which contains for each headword the years of first and last attestation and the number of quotations attested. The quotation slips from the archive form the basis of our lexicographical work, but of course we also make use of the available digital resources.

3. Digital resources

Today, DWB lexicographers use electronic databases to find additional examples or to verify their findings on a different data basis. Indeed it seems no longer possible to write a modern historical dictionary without the help of digital resources (see also Solf 2011). Ideally, these new resources should fulfill the same standards as conventional resources with regard to size and selection of texts. But they will have to fulfill a number of additional criteria if they are supposed to make the day-to-day lexicographical work easier and faster. In the following we will introduce a number of freely available digital text collections and assess them with regard to their suitability for historical lexicography according to the following six criteria: (i) size, (ii) make-up and selection of texts, (iii) verifiability of the results, (iv) quality of full texts and scans, (v) presentation of search results and (vi) search functions.

There exist a large number of collections of historical German texts. For the present purpose we will only consider *large* resources[1] that cover the time between 1700 and 1900[2]. Furthermore we will only consider resources which do not require use to consult other sources in order to verify our findings. This means that resources that do not offer full texts together with the scans of the texts on which the full texts are based will not be considered further here.[3]

Accordingly, the following three text collections will be looked at in more detail: The **Deutsches Textarchiv (DTA)**, which is funded by the Deutsche Forschungsgemeinschaft and is situated at the Berlin-Brandenburg Academy of Sciences and Humanities in Berlin. It digitizes a core inventory of German texts from different disciplines. During the first project stage (2007-2010) 650 works from between 1780 and 1900 were included. By 2014 it will comprise 1300 texts from between 1650-1780. Note that DTA is the only of the resources that can be called ‘corpus’ in a more narrow sense, i.e. there are, for example, certain criteria for the selection of texts and the texts contain linguistic annotations. The other resources are not aimed at lexicographers or linguists.

Wikisource is a multilingual online project that collects and edits texts that are not subject to copyright laws or are under a free license. For German, it covers the time from Old High German up to the 21st century.

Google Books, a service by Google Inc., contains an enormous amount of books from all languages. It has been online since 2005. Google Books aims at digitizing all books ever published. Many dictionaries now use Google Books – though often somewhat off the record –, among them the Swiss historical Dictionary *Schweizer Idiotikon* (cf. Bickel 2008), the *Deutsches Fremdwörterbuch* (cf. Brückner 2009), the French phraseological dictionary *Dictionnaire des expressions quotidiennes* (cf. Lengert 2010), the *Dictionnaire Étymologique Roman* (cf. Schweickard 2010) and the *Deutsches Wörterbuch* (cf. Solf 2011).

Table 1. Digital resources and their suitability for diachronic lexicography (accessed Oct and Nov 2011; '?' indicates that this function does not always work properly).

	Deutsches Textarchiv http://www.deutschestextarchiv.de	**Wikisource – German** http://de.wikisource.org	**Google Books** http://books.google.de
SIZE	• 532 books online from between 1778 and 1905 • 120 books are currently prepared for publication	• 23.187 texts (including poems, legends and other short texts) • New texts are added	• Overall size: 15 million books (12% of all books; cf. Bohannon 2010) • German: 110 billion words (cf. Michel et al. 2011) • Germany: books after 1871 not easily available as full texts • New books are added
SELECTION OF TEXTS	• Core inventory of German texts from different disciplines and genres • Based on DWB bibliography, complemented by scientists of different disciplines • First editions	• Users add what they like • "Authoritative editions"	• Scans the collections of academic libraries • Books that are still in print are provided by publishers
VERIFIABILITY OF RESULTS	• Full texts and scans on one page • Correct reference for every text • Stability guaranteed	• Full texts and scans on one page • References occasionally incomplete • Stability not guaranteed (but aims at presenting texts in an academically sound way)	• Full texts and scans • References sometimes wrong or incomplete • Stability not guaranteed (texts disappear, existing texts are not found etc.) • Frequency counts unreliable
QUALITY OF TEXTS	• Inclusion of texts via OCR or double-keying • Currently all full texts are corrected manually • Full texts are faithful to the original • Good quality of scans	• Good quality scans are taken over from other databases • Inclusion of texts via OCR or typewriting • Texts are corrected twice • State in editing process indicated for each text	• Inclusion of texts via OCR • Poor quality of full texts of older books • Some scans are defective

PRESENTATION OF RESULTS	• List of text passages (full text and/or scan), references • Search term is highlighted • Costumizable view: full text (original or normalized), scan, markups • Results can be sorted by date	• List of text passages (full text), references; one click is needed to get to scan and full text • Search term is highlighted (?) • Results cannot be sorted	• List of text passages (full text), references; one click is needed to get to scan • Search term is highlighted • Results can be sorted by date • Only a limited number of hits are displayed (37 pages?)
SEARCH FUNCTIONS	• Systematic access: text, author, year of publication • Full text search: phrase search, wildcards, Boolean operators etc. • Metadata search: title, author, year, period of time • Linguistic annotation (token, lemma, wordclass) finds variants (?)	• Systematic access: author, text, discipline, time of writing, place of writing, language of the original, text type, form of production • Full text search: phrase search, wildcards, Boolean operators, search for several words • No metadata search • Automatic search for similar forms (?)	• Full text search: phrase search, wildcards, Boolean operators, search for several words • Metadata search: date of publication, author, title, publisher, ISBN, ISSN • Automatic search for similar forms (?)

4. Finding examples

From a lexicographer's perspective the DTA is the standard by which the other sources have to be measured. It is superior to the others with respect to make-up and selection of texts, verifiability of the results, quality of the full texts and scans, search functions and presentation of the search results. Regarding the last point, however, further sorting possibilities, e.g. according to preceding or following word or according to word form, are desirable (however, these functions could also be taken over by an external 'example manager' tool). Unfortunately at present the DTA is too small and many less frequent words are not sufficiently attested (see table 2).

The biggest shortcoming of Wikisource is also its size (according to the numbers of hits for different searches it has around the same size as the DTA). Furthermore it provides only very restricted search possibilities for lexicographers. Although Wikisource can be accessed in different ways (systematic access, full text search), diachronic lexicographers would at least want to be able to restrict their searches to certain time spans or to sort the search results by date.

In Google Books we find a vast number of examples for virtually every word, but, as is well-known, it is defective in many ways: the full texts and the scans are sometimes of very poor quality, the verifiability of the search results cannot be guaranteed (unless we print out or download what we need), the criteria for the searches are not known, the search results are often 'noisy' because with older texts the optical character recognition method (OCR) used for digitization does sometimes not recognize words correctly (obviously this problem becomes more serious the older the texts are), frequency counts are unreliable etc. However, due to its unprecedented size Google Books cannot be neglected as a source for a historical dictionary. Obviously it can only be used for certain types of searches: the possibility to restrict searches to certain time spans allows us to search for first or last attestations of a word, for infrequent words Google Books can be used to find additional examples.

Table 2. Number of examples from the time between 1700 and 1900 for a number of compounds with *Bein* ‘leg’.[4]

Bein-	**-fraß**	**-gewand**	**-harnisch**	**-hart**	**-haus**	**-haut**	**-kleid**	**-schiene**	**-schrötig**	**-well**	**-schwarz**
DTA	1	0	(5)[5]	0	4	10	147	15	0	0	1
Wikisource	0	0	0	1	15	3	265	9	1	2	4
Google Books	6650	520	1100	3900	6400	24900	80200	1500	40	290	7780
^{2}DWB archive	13	2	5	7	36	10	178	12	3	0[6]	13

Table 2 compares the number of examples for different compounds with *Bein* ‘leg’ in the digital resources with the number of quotation slips for the relevant time span in the quotation archive. We only counted those quotation slips where the context was large enough to interpret the given headword. The figures show that for less frequent words the number of examples in the ^{2}DWB quotation archive often exceeds the number of hits for the same word in DTA and Wikisource. Note that for most headwords the archive contains further bibliographic references to examples and further references to and examples from secondary sources, which were also not included in the count.

It is to be expected that in the future there will be a corpus of historical German that is large enough for our purposes and covers all periods of German (interconnected reference corpora for Old High German, Middle High German and Early New High German are currently being compiled; there are also plans to extend DTA, cf. Geyken 2011). At present, the digital text collections can only be used to complement the existing resource, ^{2}DWB archive.

There is absolutely no doubt that such an ‘ideal’ corpus will be superior to the manual search for examples in many ways, which do not have to be mentioned here. Nevertheless we want to finish with a few – perhaps provoking – thoughts on the advantages of manually excerpted examples of historical German vs. examples drawn from digital resources (even provided that we have an ideal corpus). First, a manually compiled quotation archive is “intelligent”, i.e. the examples it contains were collected with their purpose (as examples for the dictionary) in mind. This means that the archive will contain a larger proportion of ‘good’ and usable examples than a digital corpus.[7]

Second, historical lexicographers do normally not need an unlimited amount of examples for every word. Searches for frequent words in digital corpora often return amounts of data with which it is impossible to deal. And if we restricted the search to a random sample we run the risk of losing less frequent uses or variants of a word. Other, more sophisticated filter systems (as described, for example, in Kilgarriff et al. 2008 or Didakowski et al. 2012) will not be available for earlier stages of the language in the foreseeable future. A human excerptor, on the other hand, does not write down *all* uses of a word in a given text; on the other hand it can be assumed that he or she will include new or unusual uses or meanings. We thus find that the archive contains a larger number of ‘interesting’ examples, documenting new uses or meanings (cf. also Durkin 2009, who notes that the most interesting examples for the *Oxford English Dictionary* still come from the reading programme).

Third, although tools for the lemmatization of historical texts are being developed (for German see e.g. Pilz et al. 2008 or Jurish 2010) it is doubtful whether such tools will be able to substitute skilled excerptors in tracing historical or regional variants or old or erroneous spellings – this problem will become more pronounced the further we go back in time.

5. Conclusion

We have introduced the ^{2}DWB quotation archive, which forms the basis of the lexicographical work on the big historical dictionary of German, the *Deutsches Wörterbuch* (^{2}DWB). Today, the archive is complemented by digital resources. The assessment of three freely available digital collections of text according to size, selection of texts, verifiability of search results, quality of full texts and scans, presentation of search results and possible search functions has shown that none of these resources can yet substitute ^{2}DWB archive. The DTA is superior to the other text collections with regard to most of the criteria tested. But it has to be expanded considerably if it wants to become an important source for diachronic lexicography. Until then, Google Books will remain the most important digital resource – despite its many defects. We have further argued that from the point of view of diachronic lexicography in some areas the "intelligent" examples from the quotation archive are superior to automatically retrieved examples from digital corpora.

Notes

[1] The sizes of different resources are hard to compare because they measure size differently, e.g. in words, texts or books.

[2] This time span was chosen for the sake of clarity. The corpus situation becomes more complicated the further we go back in time, but note that interconnected reference corpora for earlier periods of German are currently being compiled in Berlin, Bochum and Halle.

[3] However, at least two of them have to be mentioned here. The first is the quite well-known **Projekt Gutenberg** (http://www.projekt.gutenberg.de), which is run by volunteers and offers texts which are no longer subject to copyright restrictions. Projekt Gutenberg is not suited to lexicographers' needs. The full texts do not contain page numbers and a number of clicks are required to actually view the relevant text passages; furthermore the bibliographical references are sometimes incomplete and the texts are not always true to the original text. **Zeno.org** (http://www.zeno.org) contains texts from the 15th to the early 20th century, which are no longer subject to copyright restrictions, including many CDs and DVDs of the Digitale Bibliothek (Directmedia Publishing). In 2009 the research association **TextGrid** obtained the rights for the texts, which can now in parts be downloaded from the TextGrid website (http://www.textgrid.de). The quality of the texts in Zeno.org is much better and the references are normally correct. It can therefore be used as starting point for further research in a library or in Google Books. One further resource, which offers full texts and scans, has to be mentioned here: **Open Library** (http://openlibrary.org), which is a project of the non-profit Internet Archive and contains around 1 million books, whose full texts can be accessed via http:openlibrary.org/search/inside. Open Library seems to contain a large number of German books for the relevant time span (430 hits for *Beinhaus*, 634 for *Beinkleid*, 12 March 2012), but cannot be accessed systematically and further information about its make-up is hard to find.

[4] Searched on 20 October 2011 for the standard form. Note that this excludes one hit for *beinschrötig* in the DTA, which should have been found by such a search but was only found by searching for 'beinschr*'. The search in Google Books was restricted to 'preview' and 'full text' (the 'snippet view' is of no use for us).

[5] The examples are all from one source.

[6] But note that the quotation archive contains 15 examples from secondary sources for *Beinwell*.

[7] We are aware that such a distinction between 'good' and 'bad' examples is rather simplistic. We will have to specify what 'good' or 'bad' exactly refers to: good for inclusion in the dictionary?, good for showing the lexicographer typical uses of a word? good because a new use or meaning is illustrated? etc.

References

A. Dictionaries

Bernet, C. and P. Rézeau 2008. *On va le dire comme ça. Dictionnaire des expressions quotidiennes*. Paris: Balland.

Buchi, E. and W. Schweickard (eds.). *Dictionnaire Étymologique Roman*. (Online: http://www.atilf/DERom)

Grimm, J. and W. Grimm et al. (eds.) 1854-1961. *Deutsches Wörterbuch* (First Edition). Leipzig: Hirzel. (^{1}DWB; Online: http://woerterbuchnetz.de/DWB)

Berlin-Brandenburg and Göttingen Academies of Sciences and Humanities (eds.) 1961f. *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm. Neubearbeitung*. A-F. Leipzig, Stuttgart: Hirzel. (^{2}DWB)

Schulz, H. and O. Basler (eds.) 1990f. *Deutsches Fremdwörterbuch* 1990f. (Second Edition). Berlin: de Gruyter.

Simpson, J. (ed.). *Oxford English Dictionary* (Third Edition). Oxford: Oxford University Press. (OED; Online: http://www.oed.com)

Staub, F. and L. Tobler et al. (eds.) 1881f. *Schweizerisches Idiotikon. Wörterbuch der schweizerdeutschen Sprache*. Frauenfeld: Verlag Huber. (Online: http://www.idiotikon.ch)

B. Other literature

Bickel, H. 2008. 'Die Erschliessung neuer Kanäle: Die Volksausgabe des Idiotikons und «Idiotikon online».' Paper presented at the conference *Das Idiotikon: Schlüssel zu unserer sprachlichen Identität und mehr. Frühjahrstagung der Schweizerischen Akademie der Geistes- und Sozialwissenschaften,* Zürich, 24 April 2008, 151–162.

Bohannon, J. 2010. 'Google opens books to new cultural studies.' *Science* 17: 1600.

Brückner, D. 2009. 'Die Google Buchsuche als Hilfsmittel für die Lexikographie.' *Sprachreport* 3: 26–31.

Didakowski, J. et al. 2012. 'Automatic example sentence extraction for a contemporary German dictionary.' In this volume.

Dückert, J. 1987. *Das Grimmsche Wörterbuch. Untersuchungen zur lexikographischen Methodologie*. Stuttgart: Hirzel.

Durkin, P. 2009. 'Dialect and non-standard language in the new edition of the *Oxford English Dictionary*.' Paper presented at the *6. Arbeitstreffen der deutschsprachigen Akademiewörterbücher*, Berlin, 2-5 September 2009.

Geyken, A. 2011. 'Das Deutsche Textarchiv. Auf dem Weg zu einem Referenzcorpus der deutschen Sprache.' Paper presented at the conference *Perspektiven einer corpusbasierten historischen Linguistik und Philologie,* Berlin, 12-13 September 2011.

Jurish, B. 2010. 'More than words: using token context to improve canonicalization of historical German.' *Journal for Language Technology and Computational Linguistics* 25.1: 23–40.

Kilgarriff, A. et al. 2008. 'GDEX: Automatically finding good dictionary examples in a corpus'. In E. Bernal and J. DeCesaris (eds.), *Proceedings of the XIII Euralex International Congress: Barcelona, 15-19 July 2008*. Barcelona: L'Institut Universitari de Lingüistica Aplicada, Universitat Pompeu Fabra, 425–432.

Kirkness, A. 2012. ‘Die Lexikographie Jacob und Wilhelm Grimms im europäischen Kontext. Wörterbuchschreiber als Wörterbuchbenutzer.’ *Brüder Grimm Gedenken* 17.

Lengert, J. 2010. ‘Review of C. Bernet and P. Rézeau, On va le dire comme ça. Dictionnaire des expressions quotidiennes, Préface de Jean Artarit.’ *Zeitschrift für romanische Philologie* 126.4: 657–662.

Michel, J.-B. et al. 2011. ‘Quantitative analysis of culture using millions of digitized books.’ *Science* 14: 176–182.

Pilz, T. et al. 2008. ‘The identification of spelling variants in English and German historical texts: Manual or Automatic?’ *Literary and Linguistic Computing* 23.1: 65–72.

Solf, M. 2011. ‘Überlegungen zu den Quellen für ein historisches Wörterbuch an der Schwelle zum digitalen Zeitalter.’ Paper presented at the workshop *Perspektiven historischer Lexikographie in einem digitalen lexikalischen System*, Berlin, 28 March 2011.

Schweickard, W. 2010. ‘Die Arbeitsgrundlagen der romanischen Etymologischen Forschung: vom REW zum DÉRom.’ *Romanistik in Geschichte und Gegenwart* 16.1: 3–13.