Digital Dictionary of the 20th Century German Language

Damir Ćavar, Alexander Geyken & Gerald Neumann[1]

Berlin-Brandenburgische Akademie der Wissenschaften, DWDS
Jägerstr. 22/23, D – 10117 Berlin
{cavar|geyken|neumann}@bbaw.de

Abstract

There is no dictionary of the German language which gives a satisfactory presentation of the vocabulary of the 20th century. In this respect, German lags conspicuously behind French and English, and the cultural wealth they reflect. The DWDS seeks to make up the lost headway, drawing on the data technology of today. The outcome will not be a printed dictionary, but a digital system of dictionaries from which information can be retrieved systematically online, and specialized dictionaries generated as required.
The project is divided into two consecutive phases. The corpus will be collated with funds from the Deutsche Forschungsgemeinschaft (DFG), to be followed by lexicographical work. Both phases will be performed in collaboration with the Austrian Academy of Sciences and the Swiss Academy of Humanities and Social Sciences.

1. Aims

There is no dictionary of the German language which gives a satisfactory presentation of the vocabulary of the 20th century. Grimms' monumental opus (1854-1960, [GRI 1999]), the Dictionary of the GDR Academy (1964-1977, [WDG 1977]), and the Duden Dictionary ([DUD 2000]) hardly touch on the language of the first half of the century. If we compare this to the situation in languages such as French and English, which reflect a similar cultural wealth, it becomes highly conspicuous that German vocabulary research lags well behind. This is not only an omission on the part of researchers. It is a barrier to the employment of German as a means of communication by all who use it as their native or second language. Last but not least, it documents a failure to confront to texts of the past and to appreciate that the language of these texts is the foundation for that we use today.

A committee bringing together Hans Magnus Enzensberger, Wolfgang Frühwald, Gottfried Honnefelder, Wolf Lepenies, Christian Meier (chair), Johannes Rau, Richard von Weizsäcker and Dieter E. Zimmer will be providing ongoing support for the Academy's project.

The Digital Dictionary of the 20th Century German Language (DWDS) is an attempt to make up this lost headway step by step. It has four salient features.

First, it will draw on advanced data technology right from the outset. It follows from this that the outcome will be not simply a printed dictionary but a digital system of dictionaries from which information can be retrieved online at will and various dictionaries generated and printed as required.

Second, it will offer a user-friendly platform for different groups and a valuable aid to anyone interested in the German language. The DWDS is an academic dictionary, and it is the task of basic research to compile it. At the same time, it will be of immediate practical benefit not only to lexicographers. Last but not least, it will provide the academic basis for the development of more specialized mono- and bilingual dictionaries by other research bodies and commercial publishers.

It will not be limited to works of fiction, but treat journalistic prose and specialized texts, advertising copy and operating instructions on a par with these. It will also cover the now extensive corpuses of spoken language – everyday language, televised debates, dialect and many others. The decision to include these together with the relevant information is not just a matter of completeness; it is essential to lend the 'colloquial' 'regional' and similar labels familiar from numerous dictionaries a scientific foundation beyond lexicographers' mere intuition.

Finally, the DWDS will not be developed from A to Z like a conventional dictionary, but in a number of modules. This will permit work to be interrupted or even terminated at any time.

2. A digital dictionary

Data processing has opened up numerous opportunities for lexicographical work over the years. This potential should be used more consistently. Like any lexicographical work, the DWDS builds on a representative text corpus. In such a corpus we encounter inflected word forms. Common programs are able to locate these with ease, print them with their respective contexts, count them and sort them alphabetically. Although this information is of limited value, it is helpful for some purposes. Further benefits depend on the supplementary information linked to the actual word forms. This information includes pronunciation (RP and regional variants), lemmatization of parts of speech, valency, etc., different meanings in different contexts, first recorded use and also semantically related words.

Some of this information can be compiled quickly, even automatically. Other information requires human analysis, as it always has. But the technologies available today frequently make this a much easier task than it was in the past.

A highly comprehensive dictionary is produced very quickly, but with little analytical depth in the first instance, i.e. with only a small portion of the information described. This is a relatively simple task given the appropriate texts and a number of user-friendly utilities. After that, the analysis will proceed in the way described above. The modular structure is therefore a major

[1] The members of the DWDS project are: Manfred Bierwisch (fellow), Damir Ćavar (research assistant), Alexander Geyken (fellow), Wolfgang Klein (coordinator), Wolf-Hagen Krauth (fellow), Gerald Neumann (research assistant), Dieter Simon (fellow), Ralf Wolz (research assistant).

component of the envisaged project. The DWDS is a venture which can be developed, expanded and detailed in multiple ways, but one with a practical and academic benefit right from the outset. One aspect of its practicality is that 'paper versions' can be produced at will, printed dictionaries for special purposes, whether conventional alphabetic dictionaries, reverse dictionaries or dictionaries of synonyms.

3. Stages

The DWDS is not a long-term project in the sense that many other major lexicographical ventures are. It will be made available to all users very quickly, though admittedly with limited functions in the first instance. To gauge the practicability of the envisaged approach, the project group established by the Academy carried out a viability study in which the method and input involved in the digitization and linguistic processing of some of the sources to be included in the database (approx. 10 million words of text) were tested and a sample CD-ROM developed with this part of the corpus and a user interface based on the AltaVista search engine.

3.1. Creation of a German National Corpus

Following the successful completion of this preliminary work, the project is now in the next stage. It will last two years and aims to compile a database with a very large and balanced German corpus to be used for lexicographical purposes and covering the period from 1900 to the present[2]. Parallel databases are to be developed in close consultation with the Austrian Academy of Sciences and the Swiss Academy of Humanities and Social Sciences and will serve as the substantive foundation for a very large and balanced 'Digital Dictionary of the 20th Century German Language (DWDS)' to be produced at the Berlin-Brandenburg Academy of Sciences over the coming years in collaboration with the said academies.

The aim is to build up a corpus of 600 million running words. It would be comparatively easy to compile a corpus of several billion running words from existing electronic resources such as newspaper texts. However, it is a completely different task to compile a balanced corpus. Due to the limited financial resources, the DWDS project decided to split up the corpus into two parts: a so-called core corpus which will be a completely balanced corpus with respect to the 20th century as well as to the text genres (as described in 1.). This corpus will comprise about 100 million running words. The second sub-corpus comprises the remaining 500 million running words. It will resemble much more to an opportunistic corpus even though the DWDS tries to respect as much as possible the balance of the above- mentioned text genres and time lapse.

Descriptive markers will be added to the texts in the widespread, portable XML format. A careful structural and linguistic annotation is not only needed for querying the database, but will also facilitate the inevitable correction process during the compilation of the database.

[2] The Deutsche Forschungsgemeinschaft has kindly agreed to finance the work of the Berlin-Brandenburg Academy of Science project group on this stage of the project.

3.2. Compilation of the Corpus

The compilation of the text corpus is performed in three different steps. Texts may be available in an electronic format or only as a print version. In the latter case, these texts have to be pre-edited, i.e. selected and prepared for digitization. The digitized text is then converted into an XML-format, and processed automatically by taggers. These taggers do not only perform lemmatization of single words; they also add linguistic tags for morpho-syntactic and semantic features. After performing a few automatic quality checks the text enters a document management system (DMS). With the help of the DMS, the entire text-corpus is post-edited by human editors, who correct mistakes in the corpus, that could not be corrected by one of the automatic processing and conversion tools that were employed before. In a final phase, the corpus is prepared for search and retrieval engines, and for the use in a lexicographers' workbench.

3.2.1. Corpus Acquisition

Since many texts, especially those of the first half of the 20th century are not available in electronic format, a considerable amount of data has to be digitized. We stipulate that about one third of the texts will be available only in printed form. This means that about 35 million running words[3], or in other terms about 150 000 pages will have to be digitized, while about 65 million running words can be acquired in a digitized form.

The repartition of text genres is as follows: literature (25%), journalistic prose (25%), scientific texts (20%), and specialized texts (such as advertising texts or operating instructions, 20%), as well as spoken language (10%). All texts are selected by scientific members of the Berlin-Brandenburg Academy of Sciences – all of them being specialists of their respective subject matters -, trying to respect as much as possible the balanced and representative character of the corpus.

Cooperation with different publishers proves to be extremely fruitful for text compilation. At present, the DWDS has a cooperation with two important german editors, the *Suhrkamp Verlag* and *Die Zeit*, as well as a large public library, the *ZLB (Zentral- und Landesbibliothek Berlin)*. Further cooperation is planned with other editors. With the help of the editors, it should be possible to collect most of the literature and scientific texts. It is much more difficult to collect journalistic prose where only the last decade is electronically available. Therefore, a big effort had to be undertaken to digitize older newspaper texts. Here, the original newspapers provided by the *ZLB* have been very valuable. The same holds true for the so-called specialized texts since it is very rare to find re-editions of operating instructions or books about good behavior. Finally, the spoken language sub-corpus will consist of samples stemming from electronic archives of the *Deutsches Rundfunkarchiv* and the above-mentioned *ZLB*.

The project is aware of the German and international copyright regulations, and works out contracts with the publishers which allow the use of the respective text

[3] Due to the above-mentioned restricted resources, the DWDS project will be able to digitize only the corresponding portion of the core corpus.

corpora for lexicographical purposes and guarantee that no complete text can be extracted from the corpus.

3.2.2. Digitization and Pre-Editing

In the preliminary phase, digitization was based on scanning and OCR-processing. Depending on the input quality, correction of OCR texts can be a very time consuming task. Furthermore, conversion to XML requires in many cases manual effort. Generally speaking, training of the OCR-software, scripting for the correct association of text zones to articles, as well as automatic processing of typographic features to structural mark-up is only a viable alternative for large or at least very regular texts. However, our core corpus very often consists only of text-samples in order to respect the required diversity of our corpus.

On the other hand, we experienced that manual transcription overcomes some of these shortcomings. The correctness of OCR-software is rarely above 99%, which means that there are about 100 errors per 10000 characters. Manual transcription, if based double-keying, produces not more than 5 errors per 10000 characters for almost any input quality. Furthermore, some of the mark-up can already be done during transcription, thus making the XML-conversion an almost automatic task, a remarkable by-product of manual transcription! Therefore, we decided to cooperate with a company in China where 30 of the above-mentioned 35 million running words are presently transcribed.

It goes without saying that a careful description of the tags to be used during the transcription process is a necessary prerequisite for a successful XML conversion at a later stage. Most of the mark-up can be done by human transcribers without any knowledge of XML or any cultural background of the texts. Other mark-up, however, has to be done by native speakers.

Due to our decision to perform most of the digitization by human transcription, additional pre-editing is necessary in order to reduce as much as possible the expensive and time consuming post-editing process after digitization.

Pre-editing is performed on the basis of image scans, and commercial software is used to perform the basic operations such as copy, paste, insertion of text, etc. It consists essentially of the following steps: document selection, quality control of the input text, and mark-up of difficult parts of the document.

The selection of the document is by no means trivial since for example newspaper articles are sometimes discontinuous. Also, a preliminary quality control has to be performed at this stage since many of the older texts contain text portions which cannot be directly transcribed. Thus, it is necessary to mark-up these portions and to put them aside in order to process them, if necessary, at a later stage. Finally, difficult text portions have to be marked-up by students. The correct association of a photo with its corresponding article, or the distinction of teasers and intermediate titles, are examples for such cases which cannot be performed yet in China.

3.2.3. XML conversion

Digitization produces files in UMTF-8 format with XML mark-up. These files are validated against a DTD which varies with the text genres. The texts are then translated to a portable XML format, adhering to the standards proposed by the Text Encoding Initiative (TEI).

3.2.4. Quality Annotation

Reasonable and efficient text correction can only take place if the text is linguistically analyzed. Frequent errors are for example the confusion of abbreviations with sentence borders, the confusion of a soft-hyphen with a dash, or all kinds of misspelling. The main problem for the last error type is, that basically all proper nouns are potential candidates for misspellings.

Given the large amount of text, it would be impossible to correct each error individually in each text. It is therefore necessary to recur on error classes which enable the human corrector to work on error types rather than on individual errors.

In order to perform this correction task efficiently, the texts have to be linguistically annotated, and a document management system (DMS) has to be implemented.

3.2.5. Linguistic Annotation

Linguistic annotation consists basically of lemmatization, morpho-syntactic and semantic annotation on the word level, as well as prepositional phrase and noun phrase recognition on the phrase level. After an evaluation of several available taggers for German, we decided to use *MPRO* for this purpose. The *MPRO*-tagger was licensed from the *Institut für angewandte Informationsverarbeitung* (*IAI*) in Saarbrücken. *MPRO* is a robust chunk parser which produces a unique parse if the sentence can be parsed, or partial parses (without overlap) otherwise. It reposes on a complete morphological analyzer respecting both word-formation and composition.

Not all of the tags used by *MPRO* are also used for corpus annotation since tagging for parsing purposes is generally different from purposes for lexicographical purposes. Therefore, a specific tag set which is necessarily based on the tag set of the *MPRO*-tagger is used for the DWDS corpus.

3.2.6. Integrated workflow and DMS

In addition to linguistic annotation, a document management system (DMS) is crucial in order to perform manual correction of the above-mentioned error types efficiently.

We are currently evaluating different DMS with respect to the above-mentioned needs[4]. Technically speaking, the DMS has to be able to manage a corpus of structured and annotated texts in XML format with more than 600 million running words. It has to provide mechanisms for version control and parallel editing of one document. For the editing process, a retrieval and search mechanism is a necessary component of the DMS.

4. Conclusion and future perspectives

The idea of a Digital Dictionary of the 20th Century German Language (DWDS) reposes on two basic ideas: New techniques of data processing have to be used right from the outset in order to create a balanced and sufficiently large corpus. This corpus is not only the basis for future lexicographical analysis, but has an independent

[4] It is not clear yet, which system will be used for this stage. There are basically two possibilities: either a use of a XML-compliant retrieval engine such as Verity, or the use of a database system (DB2 by IBM, or Oracle), in combination with own tools that have to be developed.

value as a German National Corpus of the German language of the 20th century.

Apart from building the corpus, a lexicographers' workbench will be implemented that will form the focal element of the lexicographic project stage starting in 2002.

5. Bibliography

[DUD 2000] Das große Wörterbuch der deutschen Sprache, Mannheim, Leipzig, Wien, Zürich: Dudenverlag.

[GRI 1999] Grimm, Jacob; Grimm, Wilhelm, Deutsches Wörterbuch, München: Deutscher Taschenbuch Verlag (Lizenzausgabe).

[WDG 1977] Klappenbach, Ruth; Steinitz, Wolfgang, Wörterbuch der deutschen Gegenwartssprache, 8. bearbeitete Auflage, Berlin: Akademie Verlag.