The DWDS corpus: A reference corpus for the German language of the 20th century

Alexander Geyken – Jan. 2006

D R A F T – please do not circulate

To appear in *Fellbaum, Christiane* (Ed.) : *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London (Continuum Press)

Abstract

The DWDS corpus, constructed at the Berlin-Brandenburg Academy of Sciences (BBAW) between 2000 and 2003, consists altogether of over a billion words of running text. Corpus building continues to be an activity at BBAW. The current corpus consists of two parts: a core corpus and an extended corpus. The core corpus contains approximately 100 million running words, balanced chronologically and by text genre in approximately 80,000 documents. About 40 percent of these, i.e. approximately 160,000 pages, were digitized from printed resources by the project. The remaining texts were obtained from publishing houses or donated by contributors. The core corpus is unique in German speaking countries and constitutes, for German, a resource equivalent in quality to the British National Corpus. The extended corpus contains more than 900 million text words. It is an opportunistic corpus, consisting essentially of newspaper sources from the last 15 years. Copyright clearance has been obtained from major publishing houses, enabling DWDS users to access the works of important literary and scientific authors including Heinrich Böll, Jürgen Habermas, Victor Klemperer, Karl Kraus, Siegfried Lenz, and Thomas and Heinrich Mann. All the texts of the core corpus are lemmatized and part-of-speech tagged, and can be queried with DDC (Dialing DWDS Concordancer), a linguistic search engine, on the project's web site. It is intended to gradually add further texts of the 20th century and to extend the corpora to the 21st century and also to texts before 1900.

1. Introduction

Over the course of two and a half years between 2000 and 2003, two corpora of 20th/21st century German language were compiled at the Berlin-Brandenburg Academy of Sciences (BBAW).[1] The first is a balanced core corpus ("Kerncorpus"), a reference corpus for the German language of the 20th century; the second is an opportunistic supplementary corpus ("Ergänzungscorpus") (Geyken and Klein 2004). Both corpora represent the first step towards a larger resource, the *Digital Dictionary of the 20th/21st Century German Language* (DWDS). Like other dictionary projects, the DWDS project falls into two phases: a corpus compilation phase, followed by the lexicographic work itself. In order to understand the design principles of the DWDS corpora it will be useful to present here the main ideas of the project.

There were three main motivations for the project.: First, there is currently no dictionary of the German language which offers a satisfactory representation of the lexicon of the entire 20th century: Grimms' monumental opus (1854–1960), the Dictionary of the GDR Academy (1964–1977), and the Duden Dictionary hardly touch on the language of the first half of the century. In particular, the vocabulary of the second half of the German Empire of 1871–1918, the Weimar Republic, and the Third Reich are not reflected. Moreover, recent developments in the current language need to be objectively studied and recorded in

greater detail than is presently possible. These facts point not only to a failure to confront the texts of the past, they also indicate the presence of a barrier to the full understanding of German as a means of communication by all who use it as their native or second language (Schmidt 1997).

Second, traditional print dictionaries are compiled in an alphabetical order and structured into individual and discrete entries. This has some serious disadvantages for the consistent description of words belonging to one and the same lexical field, as the initial letters of these words are likely to be distributed across the alphabet (cf. e.g. Harm 2005, Schulz 2001). To show meaningful relatedness, a new dictionary should not order its words alphabetically but rather by lexical categories, types of syntactic constructions, and lexical fields.

Third, even though all of the aforementioned dictionaries have invested an enormous amount of work and skill in the compilation of large databases of example sentences, they do not form a balanced corpus of German. They either rely on manual excerption of words and word uses from texts, as in the case of Grimm's dictionary, or they use a mixture of manual excerption and automatically compiled electronic opportunistic corpora—mostly newspaper corpora—as in the case of Duden. Manual excerption requires a large number of persons (usually not lexicographers) to transcribe unusual or "interesting" words or word uses from various resources; in the first edition of Grimm, for example, the excerpts are based on approximately 25,000 different sources. The excerpts are stored on paper cards and files and, more recently, in electronic databases. Since the excerptors typically neglect high frequent words or "trivial" word uses, the number of examples remains relatively small for all words and does not reflect their actual frequency in texts. Compared to electronic corpora in which complete texts are stored, manual excerpts have two disadvantages: the excerptors may inadvertently overlook important words or word senses while they read a text. There is also the danger that the excerpts lack consistency; each excerptor may have different criteria for what he selects, and these criteria may change over time. Electronic corpora avoid these problems. The challenge here is to filter out interesting words and word senses in a large mass of data. Recently developed lexicographers' workbenches have begun to demonstrate the efficiency of tools for the compilation of large new monolingual dictionaries (e.g. Kilgarriff 2002, Heid 2004). Apart from achieving better consistency and completeness of examples there is a third factor that pleads for the use of large and balanced corpora—frequency. The frequency of examples in corpora can reveal to what extent particular word forms and senses or multi-word expressions are typical, i.e., we can determine their relative frequency in actual German texts.

2. The need for a new corpus

At the outset of the DWDS project at the BBAW in 1999, no satisfactory corpora of 20th century German existed. On the one hand, there was the LIMAS Corpus, a first-generation corpus, created in 1973, which followed the model of the Brown corpus (Kučera and Francis 1967). The LIMAS Corpus consists of 500 text samples of 2000 tokens each from the year 1964, with more than 20 different text genres. Their distribution was intended to reflect proportionally the relative importance of each genre. Even though the LIMAS corpus is generally considered a balanced corpus, with its 1 million tokens and approximately 100,000 types it is far too small to constitute the text basis for a large monolingual dictionary. On the other hand, very large text collections are available at the Institut für deutsche Sprache (IDS) in Mannheim (www.ids-mannheim.de). These texts currently comprise more than 2 billion written tokens and 4400 hours of recordings of

spoken language. However, the written corpora focus mainly on recent newspaper texts, in particular many selections from provincial newspapers. Life writing texts (such as autobiographies or letters), specialised texts (advisory books, instruction manuals, advertisements), scientific texts and to a certain extent also national newspapers, nonfiction, and substantial prose works are underrepresented in this corpus, at least in the publicly available fragment. Moreover, the IDS corpus contains very few texts from the first half of the 20th century and is thus not chronologically balanced.

Due to the lack of large and balanced German corpora, new resources provided by computational linguists as well as new dictionary editions mostly rely on a few available newspapers. For example, German tree banks like Negra (Skut et al. 1997) or TüBa-D/Z (Telljohann et al. 2004) rely on the *Frankfurter Rundschau* and *die tageszeitung (taz)*, respectively. Likewise, the corpora underlying the lexicographers' workbench created for LexiView, a joint project between the large publisher Langenscheidt and the University of Stuttgart (Heid 2000), though undoubtedly a step in the right direction, are opportunistic and insufficient: "Corpus material has been taken from freely available or specifically licensed newspaper texts, among others *Frankfurter Rundschau*, *Stuttgarter Zeitung* 1992/93, a total of 350 million words. Our corpus is not balanced, as a general balanced corpus of German is only being created, e.g. at BBAW" (Heid 2004).

More recently, general-language corpora of German were created on the basis of Web resources. For example, Sharoff (Sharoff 2004) built general-language corpora of English, Russian, Chinese, French, Italian and German, all similar in size to the BNC, on the basis of web resources. More than 41,000 URLs contributed to the German Internet Corpus. Even though this corpus provides interesting material, it does not obviate the need for a reference corpus as we define it below.

3. Corpus design requirements

Corpora are compiled for a variety of different purposes, e.g., as an empirical basis for NLP programs, as reference materials for the compilation of grammars or dictionaries, as the basis for language acquisition, or for the study of language history. Corpora are generally multi-purpose, even though a single corpus can never be used as a basis for all goals. For example, the British National Corpus (BNC) is considered to be a general-purpose, general-language corpus, designed to be used for NLP as well as grammar studies or as the basis for large monolingual dictionaries. However, one would hardly use the BNC for conducting studies of language acquisition or for diachronic investigation.

The main purpose of the DWDS corpus is to serve as the empirical basis of a large monolingual dictionary of the 20th/21st century. The ambition of this dictionary is to offer more subtle linguistic descriptions of the semantics and syntagmatics of lexical items than was possible before the availability of such a large body of evidence. Hence a desideratum for the DWDS corpus is that it should be "representative" of the German language, although representativeness is admittedly a problematic concept in corpus linguistics. Indeed, there is widespread agreement among corpus linguists that, strictly speaking, representativeness in a statistical sense cannot be obtained for corpora, because of the difficulties associated with defining the underlying population. To mention just one example, there is no generally accepted definition of "all text types of a language" (see Bergenholtz 1990 and Biber 1994). There are also practical obstacles: too many German texts of the 20th century are not yet digitized; the situation is even worse for the spoken language. Therefore, many corpus linguists have abandoned the notion of

representativeness and replaced it by the more modest notion of "balance" (Kilgarriff & Grefenstette 2003).

In view of these observations, the first desideratum for the DWDS Kerncorpus is reformulated: instead of being representative, it has to be balanced with respect to text types. Examples of balanced corpora are the above mentioned Brown corpus and the Limas corpus. Those corpora are however not sufficiently large to form the basis of a large monolingual dictionary since they contain only 1 million tokens and approximately 100,000 types. Thus, in addition to being balanced the DWDS Kerncorpus must also satisfy the criterion of size: it must be large enough for its purpose. A third condition to be met is that the DWDS Kerncorpus must contain a considerable amount of influential and important literature. This requirement goes back at least to Samuel Johnson, who stated in his preface to his *Dictionary of the English Language* that he selected the examples wherever possible from 'writers who were masters of elegance or models of style' (Hanks 2005: 264). For 20th century German, major writers represented in the DWDS Kerncorpus include Thomas Mann, Franz Kafka, Alfred Döblin, Friedrich Dürrenmatt and Robert Musil. Texts by Albert Einstein, Walter Benjamin, Max Weber and Jürgen Habermas should be comprised as well as the writings of Kurt Tucholsky, Alfred Kerr, and Karl Kraus.
To sum up: the three desiderata for the DWDS Kerncorpus are to create a large balanced corpus containing a considerable number of influential writings and writers. According to Sinclair (Sinclair 1994), these properties characterize a reference corpus, i.e., a corpus that can be used as a basis for "reliable grammars, dictionaries, thesauri and other language reference materials".

4. Design of the DWDS corpus

In accordance with the desiderata stated above, the DWDS project created the DWDS Kerncorpus, a reference corpus of the 20th century German language. The DWDS Kerncorpus consists of 100 million tokens, thus matching in size the British National Corpus. It is a balanced corpus of German texts of the 20th century, i.e., it is roughly equally distributed over time and over five genres: journalism (approx. 27% of the corpus), literary texts (26%), scientific literature (approx. 22%) and other nonfiction (approx. 20%), as well as a smaller number of transcripts of spoken language (5%). Besides journalistic texts (newspaper reports and articles from periodicals taken from more than 50 different newspapers and magazines), it contains literary monographs, poetry and dramatic works from major German writers. Additionally, popular works of light fiction are included. Nonfiction texts such as cookbooks, maintenance manuals, and guides to etiquette are included, as well as important scientific texts. The DWDS Kerncorpus also covers to a certain extent corpora of spoken language (see below).

In addition to the Kerncorpus, the DWDS project also compiled a much larger corpus from electronic versions of daily and weekly newspapers of the 1990s, such as *Frankfurter Allgemeine Zeitung (1994–2000)*, *Frankfurter Rundschau (1990–2000)*, *Neue Zürcher Zeitung (1993–2000)*, *Spiegel (extracts between 1990 and 2000)*, *Die ZEIT (1996–2000)*, *taz (1986–2000)*, and *konkret (1980–2000)*. This opportunistic corpus, the DWDS Ergänzungscorpus (supplementary corpus), comprises approximately 900 million tokens gathered in two million articles. Both the DWDS Kerncorpus and the DWDS Ergänzungscorpus are used in the Wolfgang-Paul-Preis project for the linguistic description of selected German verb-noun idioms (Fellbaum 2004).

The major goal of the DWDS project has been to create the DWDS Kerncorpus, which satisfies the criteria described earlier. Before discussing the text selection and the digitization steps in more detail, we address two questions that might arise. One, why was the period between 1900 and 2000 chosen? Two, why is the corpus restricted to only five genres? Concerning the first question, the 20th century is arguably a somewhat arbitrary time barrier that does not constitute a clear-cut historical time period. Why did we not choose the beginning of the 1st World War, or its end? Wouldn't 1870, 1933, or 1945 mark more interesting historical boundaries? These dates could be justified, but they are not uncontroversial either. We decided on a clearly marked time period, the entire 20th century, without excluding the possibility of extending the corpus back to 1870 or even beyond. Concerning the second question—why distinguish only five genres and not a more fine-grained or richer classification—we were guided by the practical consideration that fewer genre distinctions make the daily corpus work easier. Again, we do not exclude the possibility of enhancing the text classification. As more and more texts are encoded following the TEI-guidelines, it would be comparatively straightforward to introduce more finely graded text types or to classify the texts according to a different code.

The compilation of the DWDS Kerncorpus is proceeding in four steps, which are explained in more detail in the following sections: text selection, copyright acknowledgements, digitization and conversion to a structured format, and text sampling.

5. Text selection

We describe the text selection procedures for each subcorpus. The complete bibliography of the current DWDS Kerncorpus can be found on a database on the website (www.dwds.de). The corpus is continually extended. The greatest problem for its extension is neither text selection nor digitization into a structured format, but the administrative task of negotiation copyright clearance (see below).

A. Prose, verse and drama (26% of the DWDS Kerncorpus)

For every year between 1900 and 1999 three longer prose works were selected: two longer ones that are "classical" literary works and one of light fiction. Obviously, there is a smooth transition between the two genres, but the purpose of this partition is to provide a certain balance, since from the point of view of the influence on the development of the German language, "light fiction" authors like Heinz Konsalik and Mario Simmel may turn out to be as influential (given their wide readership) as more acclaimed writers such as Thomas Mann and Günter Grass. Additionally, small samples of science fiction (e.g. 'Perry Rhodan'), detective stories (e.g. 'Jerry Cotton'), adventure (e.g. Fritz Steuben), and love stories were included. Furthermore, we selected some 20 children's books (e.g. Ottfried Preussler).

The basis for the selection of plays was 'Reclams Schauspielführer' (Reclam's Guide to the Theater) (Stuttgart 1996). The selection of poetry is particularly subjective, the basis for the text selection here was the anthology by Karl-Otto Conrady 'Das große deutsche Gedichtbuch' (The Big Book of German Poetry) (Frankfurt 1977).

Selection proceeded as follows: the project team established a provisional list based on official lists as well as on the expertise of the project members. Members of the Academy of Science were then asked to comment on that list, among them three specialists in German studies (Conrad Wiedemann, Wolfgang Frühwald and Wilhelm Vosskamp). They were asked to nominate for each year those texts that they consider as being the most important

and influential, both from the point of view of the German language and in terms of their popularity. They were also given the option of suggesting new titles.

For example, the following literary texts were chosen for the year 1953:

-Wolfgang Koeppen, *Das Treibhaus*, a testimonial of the Bonn Republic in the post-war years.

-Friedrich Dürrenmatt, *Der Verdacht*. A famous detective story, in which the author describes the final challenge facing his character chief inspector Hans Bärlach.

-Albert Vigoleis Thelen, *Die Insel des zweiten Gesichts*. In this book Thelen gives an account of his life in Mallorca from 1931–1936.

B. Newspapers (27%)

Newspaper reports and articles from periodicals were selected from more than 50 different national and regional newspapers and magazines.

From 1900–1933, newspaper samples were taken in regular intervals from Berlin (*Berliner Tageblatt* and *Vossische Zeitung*), Frankfurt (*Frankfurter Zeitung*), Cologne (*Kölner Zeitung*, 1900–1922) and Munich (*Münchener Neueste Nachrichten*). Additionally, small samples of regional newspapers such as the *Naumburger Tageblatt* and the *Aschaffenburger Nachrichten* were selected.

From 1933–1945, the *Völkische Beobachter* was contrasted with several hundred articles from various dissident and exile newspapers: *Das andere Deutschland* (1938–1939) published in Buenos Aires, *Der deutsche Schriftsteller* (1934–1937), the *Pariser Tageblatt* (1933–1936) and *Pariser Tageszeitung* (1936–1940), *Neuervorwärts* (1933–1940), edited by Kurt Schumacher, and *Die Zeitung* (1941–1945) published in London.

From 1945–2000, regular newspaper samples from Berlin (*Berliner Tagesspiegel*), Frankfurt (*Frankfurter Allgemeine Zeitung*) and Munich (*Süddeutsche Zeitung*) were taken. Additionally, smaller samples from GDR newspapers were selected: *Berliner Zeitung* and *Neues Deutschland*.

In addition to these periodic samples, samples reporting on specific events were collected. For the year 1900, the sampled event was the World Fair in Paris (Nov. 12th), for 1901 it was the first Nobel Prize award ceremony (Dec. 10th), for 1902 it was the end of the Boer War (May 31st). The idea behind this selection was that certain words or expressions were coined or given currency in connection with these historical events. The comparatively small list of events was complemented by large samples of Keesing's "Archiv der Gegenwart" (Archive of the Present Time) (AdG). The AdG summarized on an almost daily basis from 1931 until its end in 2004 the main events reported by news agencies, daily newspapers, and magazines.

C. Science (22%)

More than 100 members of the Academy of Sciences were asked by the DWDS project team to list, for each decade, up to four works that they consider to be the most important for their respective disciplines. The academy members represent all major scientific disciplines and fields of knowledge. The results of this survey constitute the basis for the science corpus. One result of the responses was that since 1980 almost all major scientific publications by German scientists, with a few exceptions, were published in English.
In particular, we selected for each year:
(a) on average one important scientific monograph. Here too, we tried to find a reasonable balance among the different disciplines.

(b) on average four articles from scientific journals, in a yearly rotation between scientific disciplines. In particular, we referred here to the journal "Forschungen & Fortschritte" (Research and Progress), a journal that was published at the Academy of Sciences between 1925 and 1944. Many important authors were published in that journal, including for example Wolfgang Köhler, Albert Einstein, Max Planck and Friedrich Maurer. For the period after 1980 we selected samples from two popular scientific journals, the journal *Bild der Wissenschaft* as well as the magazine *Peter Moosleitner*.

D. Other nonfiction (20%)

This subcorpus comprises self-help literature such as car repair manuals, cookbooks, guides to etiquette, and law texts. Moreover, it contains texts rarely considered in lexicography: user manuals, prescription drug information, theater and concert programs, and advertisement texts. These subgenres have had a considerable influence on the present-day language.
For each decade, the following texts were chosen: a cookbook, a popular healthcare book, a travel guide, a guide to etiquette, smaller samples of technical documentation and user manuals, several drug information sheets, and various advertisement texts. Also, large samples of legal texts were taken from the two collections "Schönfelder" and "Sartorius". As with the subcorpus of journalistic prose, attention was paid to balance texts from West Germany with corresponding works in East Germany.

E. Transcriptions of spoken language

Hardly any speech recordings older than 40 to 50 years are available. For the last two decades, corpora are available containing everyday conversations, radio and television interviews, and recordings of dialectal speech. However, the resources needed for the transcription of large amounts of unscripted conversation are still substantial. Given our time and budget constraints, we decided to collect transcriptions of non-spontaneous speech. In particular, the DWDS Kerncorpus contains 200 samples of radio interviews from the period before 1945 that were transcribed in cooperation with the Deutsches Rundfunkarchiv (German Radio Archive). For the period after 1945, we chose transcripts of German and Austrian parliamentary debates, transcripts of conversations with emigrants to Israel, transcripts of the TV debate *Literarisches Quartett*, and radio features from Deutschlandfunk (German Radio).

Texts from Austria and Switzerland

Texts from Austria and Switzerland are deliberately underrepresented in the current corpus because plans for the co-operation and coordination of corpus compilation were made with the Austrian Academy of Sciences and the Swiss Academy of Humanities and Social Sciences.

6. Copyright issues

The text selection was conducted independently of the copyright status. Since most of the important texts of the 20th century are still copyrighted one major task of the project has been to convince authors and copyright owners—in most cases via the publishing houses—to collaborate in the compilation of the DWDS Kerncorpus. At the beginning of the project, a committee was formed of prominent public personalities, including Hans Magnus Enzensberger, Wolfgang Frühwald, Gottfried Honnefelder, Wolf Lepenies, Christian Meier,

Johannes Rau, Richard von Weizsäcker and Dieter E. Zimmer. This committee not only advised the project but in particular lent authority to the project's negotiations with publishing houses.

The acquisition of copyright protected texts for the DWDS Kerncorpus takes place at several levels. A minimum prerequisite for the compilation of the corpus is to obtain the permission to use the texts project-internally for lexicographic work. Our desideratum, however, was more ambitious: the DWDS Kerncorpus at least should also be publicly available as a resource. Nevertheless, it was clear from the outset that publication in the same manner as the Brown, Limas or the BNC corpora was ruled out: no publishing house would give away rights for entire prose works of, for example, Günter Grass, Hermann Hesse, or Thomas Mann—Nobel-Prize winners whose works continue to be widely sold. A compromise had to be found to convince publishing houses to grant the rights to the DWDS project: on the one hand, the publication of the Kerncorpus via a web interface would provide a valuable resource for a wide spectrum of linguistic and lexicographic research; on the other hand, the project has to ensure that no copyright is violated. In particular, copying of entire texts via a corpus query must be prevented. We implemented several technical devices as a basis for negotiations with publishers. Depending on the copyright acknowledgements one or more of the following procedures has been applied:

- A sampling procedure guaranteeing that the DWDS Kerncorpus contains only selected samples, not the entire text.
- A flexible mechanism to display query results with variable context windows. Depending on the agreement reflected in the acknowledgements, the context window of a query hit varies between 7 words (minimal citation context) to 1 or 3 sentences or even a paragraph.
- A password protection for copyrighted texts. In addition, all users have to agree to use the texts only for non-commercial use.
- Anonymization of named entities (for example this can be particularly important for private letters).

A third aspect of copyright protection concerns the fact that the copyright holders for the electronic rights are in some cases not the publisher, but the author; this is often the case with older texts, where electronic use of a text was not yet a known application. Generally, the publishers prefer in such cases to ask the author for permission, a procedure which is always feasible, as when a newspaper publisher with thousands of authors would have had to contact all authors individually, or when there is a dispute among a writer's heirs.

Given these difficulties, it is not surprising that negotiations with publishers are slow and time-consuming. A first breakthrough for our project was the agreement reached with Suhrkamp, a publishing house which is not only famous for its many important literary and scientific authors but also known for its restrictive policy to disseminate texts freely. The agreement with Suhrkamp granted permission to use the texts of 22 authors including Uwe Johnson, Hans Magnus Enzensberger, and Martin Walser, as well as Jürgen Habermas, Theodor Adorno, Walter Benjamin, and Ernst Bloch. Other publishing houses followed: Aufbau, Diogenes Verlag, Eichborn, S. Fischer Verlagsgruppe, Hoffmann & Campe, Kiepenheuer & Witsch, K.G. Saur Verlag, Spiegel, Suhrkamp, Ullstein, ZEIT, Zsolnay as well as Deutsches Rundfunkarchiv and Directmedia. At present (January 2006) the project has obtained the permission of 15 publishing houses to make more than 71% of the texts of the DWDS Kerncorpus publicly available via the DWDS web site. The remaining 29% percent are available only for internal use (see below, section on corpus query).

7. Digitization

As mentioned, the initial text selection was carried out without considering the copyright status and the availability of the texts in electronic format. It turned out that approximately 60% of the selected texts of the Kerncorpus were already available in electronic format. These texts were either purchased as CD-ROMs or acquired directly from the publishing houses, and the task here was limited to the conversion of the data into a structured format (see the section on annotation below). The rest, i.e. 40 million tokens, corresponding to roughly 160,000 pages, had to be digitized from the printed format.
There are two different methods for full text digitization: Optical character recognition (OCR) or manual transcription. For many applications, OCR processing is the preferred alternative because of its cost-effectiveness. A recognition rate of 95% to 99% is considered acceptable, and corrections, if carried out at all, are restricted to named entities, dates and events in order to create a key word index. The situation is different for lexicographic purposes. Here, potentially all words are key words. Hence, the error rate needs to be very low. 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, a process including the 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 option for large or at least fairly regular texts. However, the DWDS Kerncorpus very often includes only smaller text samples in order to achieve text diversity.

Manual transcription, even though it is more expensive, overcomes some of these problems. Even when the original is of very good quality, the accuracy of OCR software is rarely above 99%, which means that there are about 100 errors per 10,000 characters. Manual transcription, if based on double-keying, produces not more than 5 errors per 10,000 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 completely automatic task, a remarkable by-product of manual transcription! These considerations led us to cooperate with a company in the People's Republic of China, where 35 million of the above-mentioned 40 million tokens were transcribed. The remaining 5 million tokens—generally regular texts of higher print quality—were digitized with OCR software.

It goes without saying that a careful description of the tags used during the transcription process is a necessary prerequisite for successful XML conversion at a later stage. Most of the mark-up can be done by human transcribers without any knowledge of XML or of the contents of the texts. Other mark-up, however, has to be done by native speakers (see below). Since most of the digitization is done by human transcription, additional pre-editing is necessary to reduce as much as possible the expensive and time-consuming post-editing process after digitization.

Pre-editing is done on the basis of image scans, and commercial software is used to perform basic operations such as copy, paste, insertion of text, etc. Pre-editing consists of the following steps: document selection, quality control of the input text, and mark-up of difficult parts of the document.
Document selection is by no means trivial, since for example newspaper articles are sometimes discontinuous. Older newspapers in particular pose problems for automatic clipping routines. Also, a preliminary quality control has to be performed at this stage, since many of the older texts contain segments that have suffered water damage or are of poor microfiche quality and cannot be transcribed straightforwardly. It is necessary to mark up

these portions and to put them aside for possible later processing. Finally, certain text segments must be marked up by people who know the language: examples for that are the correct association of a photograph with the corresponding article, or the distinction of teaser titles and intermediate subtitles.

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 transformed into the final xml format, adhering to the guidelines of the Text Encoding Initiative (TEI).

8. Structural Annotation

As mentioned, the structural annotation of the texts of the DWDS Kerncorpus follows the TEI guidelines. The massive digitization task required that 160,000 pages of heterogeneous texts had to be annotated and more than 150,000 documents from many different origins had to be converted, and a reasonable compromise had to be found between the depth of mark-up and the available budget. The following information was encoded: page-breaks, footnotes, titles, chapters (up to 12 division levels), paragraphs as well as prefaces and epilogues. Other lexicographically less relevant information such as registers and indexes were not transcribed. Formulae and tables (if they contained numbers) were marked up with begin and end tags, but their content was not transcribed. Furthermore, column breaks and line breaks were not annotated except in poems, where lines and stanzas are basic encoding elements. For plays, the lines, the characters' names, and the stage directions were encoded. Similarly, utterances and speakers' names were encoded for interviews. A start has been made on normalizing the orthography, using the element <orig> recommended by the TEI guidelines: here the original orthography is stored as an attribute, while the normalized variant corresponds to the element content. In this way the original orthography is preserved for token search and for the display of the KWIC lines, while at the same time, morphology tools can use the normalized variant for lemmatization. Hyphenated words at the end of lines are preserved in the XML files and their status as soft-hyphen, hyphen or dash is resolved during linguistic processing. Finally, font changes (e.g. from gothic to antiqua) are annotated as well as font variants like italics, boldface, spaced or underscore.

In addition, the guidelines for TEI headers were applied to all texts. Here, the following information was encoded: first date of publication, copyright status, genre, and bibliographic reference. If the original publication was not available or if a good but later electronic edition was available, bibliographic references to both are given. This is the case for all electronic editions of DirectMedia (well known for their Digitale Bibliothek), where the file origin is a CD and not the original print version. Page numbering follows the CD in this case. Another case are speeches, where the file origin is not the speech itself but a transcription that was made later. Here, it is important to respect the correct bibliography as well as the original date of publication.

The copyright status corresponds to the copyright acknowledgement agreed to between the publishing houses and the DWDS project. It contains information about whether the entire text or only parts of it may be shown and how much context may be displayed. It is indicated whether the copyright holder does not correspond to the original publisher. This is for example the case for Remarque's book "Im Westen nichts Neues", where the copyright owner (Kiepenheuer & Witsch) differs from the original publishing house (Propyläen).

9. Linguistic Annotation

All the texts of the DWDS Kerncorpus have been annotated using the TAGH morphology which is presented in more detail in (Geyken & Hanneforth 2006). We focus here on its main aspects. TAGH is a system for automatic morphological analysis of German word forms. It is based on a stem lexicon with allomorphs and a concatenative mechanism for inflection, derivation and composition. Weighted finite-state automata (FSA) and a "cost function" are used in order to determine the correct lemmatization of complex forms: the correct segmentation for a given compound is intended to be the one with the least "cost". Thus, an analysis with fewer segmentations is preferred over one with more segmentations. This simple mechanism rules out wrong analyses remarkably well. For example, the best analysis for the compound *Schadstoffanreicherung* (accumulation of toxic substances) is the semantically plausible SCHADSTOFF#ANREICHER~UNG but not SCHAD#STOFF#ANREICHER~UNG since the second segmentation incorrectly decomposes the lexicalized compound Schadstoff (toxic substance) into a verb-noun compound schad~en (to damage) and Stoff (substance). Likewise this mechanism prefers for the plural *Abteilungen* (departments) the correct lemma *Abteilung* (department) instead of the wrong ABTEI#LUNGE (ABBEY#LUNG). Of course this mechanism is heavily dependant on the quality of the stem lexicon which must be as complete as possible with respect to semantically opaque compounds.

TAGH is based on a large stem lexicon of about 80,000 stems and an additional 200,000 named entities that are in common use. The lexicon was compiled over a period of more than five years on the basis of a large selection of newspaper corpora and literary texts. The number of analyzable word forms is increased considerably by more than 1000 different rules for derivational and compositional word formation. The recognition rate of TAGH is more than 99% for modern newspaper texts and about 98.3% for the DWDS Kerncorpus. The lower recognition rate in the latter case is mainly due to unrecognized named entities, non-standard regional variants (such as *Ick*, *nit*, *wa* etc.), typing errors, foreign words (*mon*, *the, que*), non-standard abbreviations (*stellvertr.*, *Kammerorch.*) and words spelled according to the historical orthography of before 1902—which in practice did not disappear until almost two decades later.

In addition to lemmatization, the texts are annotated with lexical categories according to the STTS tagset (www.sfs.uni-tuebingen.de/Elwis/stts/). For the disambiguation of the lexical categories, a part-of-speech tagger called *Moot* is employed (Jurish 2004). *Moot* is a statistical tagger that disambiguates lexical classes. A lexical class can be any subset of the set of STTS tags. In addition to the trigram-based routines, the system considers user-defined a-priori sets of possible analyses (so called lexical classes) for each input token. By this means it is possible to restrict the analysis suggested by the tagger to the proposed classes by the morphology—in our case the TAGH morphology. This has the advantage that the tagger chooses only among the morphologically plausible classes and not among the entire tagset. In comparison with a traditional hidden Markov model that does not make use of prior linguistic knowledge, the method adopted by Moot leads to a reduced tag association error rate of up to 21%.

10. Sampling the DWDS Kerncorpus

In December 2005, the total text base for the DWDS Kerncorpus comprised 272,215 documents with 254,293,835 tokens. From this text base a balanced text corpus of about 100,000,000 tokens is extracted with a sampling procedure (s. below). The entire text basis

itself has a heavy bias towards newspapers such as, for example, the *Archiv der Gegenwart* with more than 80,000 articles. Moreover, the text base contains CDs consisting of collected works such as for example Victor Klemperer and Kurt Tucholsky, or entire proceedings, among them are those of the Nürnberg war trials in 1945 (73 days of legal proceedings). It is clear that a selection of these texts must be made to achieve balance. The sampling procedure respects the constraints for the DWDS Kerncorpus, i.e. it builds the maximally balanced corpus out of a total text base that is distributed equally over the entire 20th century. The balance of the selection follows the distribution of the five text types in the Kerncorpus, i.e. 27% for journalistic prose, 26% literature, 22% scientific texts, 20% other nonfiction and 5% transcriptions of speech. In order to determine the maximally balanced corpus, we identify the decade with the least number of tokens. The sampling process then calculates the expectation value according to the text type distribution. For practical reasons, this algorithm cannot be applied strictly since the decades are not equal over the 20th century, e.g. the 1940s had a much sparser text production than the 1920s. Therefore, we assume that text-type decades can deviate from the mean value. The deviation depends on the difference between the minimally balanced corpus and the corpus size to be attained. Moreover, the sampling procedure begins with a so-called 'initial corpus'. The initial corpus consists mainly of texts by major writers and scientists as well as of texts that are considered to be of a high lexicographic interest. This guarantees that works by Thomas Mann, Franz Kafka, Günter Grass, Albert Einstein, Max Planck but also of Heinz Konsalik and Walter Moers will be present independently of the sampling procedure, whereas for example newspaper articles by Kurt Tucholsky or letters by Wilhelm Busch are selected randomly. The start corpus consists of currently 22,023 texts. Of course, the start corpus must not exceed the boundaries of the required text distribution, i.e. we had to ensure beforehand that no decade and no text class in the start corpus exceeds the required token number.

The current DWDS Kerncorpus (version 0.95) consists of 79,322 documents. The corpus comprises 100,600,993 tokens (122,816,010 including punctuation marks and numbers) and 2,224,542 types.

While the DWDS Kerncorpus is balanced with respect to text types and decades (over the 20th century), a deviation of 10% is admitted for each decade/text type combination (s. appendix 1 for the detailed distribution). There are currently two exceptions to this. This concerns the 1950s and 1960s of the subcorpus of other nonfiction texts. They consist only of 1,29 million resp. 1,36 million tokens which is 35% resp. 30% less than the required mean of 2 million tokens of this text. Since it is quite unlikely that electronic re-editions of manuals or self-help books will be published, we currently digitize another 1,35 million tokens, which will be integrated in the next version of the DWDS Kerncorpus. Furthermore, the speech subcorpus has not yet been fully compiled. Here, approximately 500,000 tokens are still missing. We plan to add them in the coming nine months. Thus, a corpus of transcripts of spoken language will be part of the next version of the DWDS Kerncorpus in the fall of 2006.

11. Querying the DWDS Kerncorpus

The DWDS Kerncorpus is publicly available from the project's website www.dwds.de. All versions of the corpora are preserved. Thus, users will have access not only to the current but also to older versions of the Kerncorpus.

The basic tool for searching the DWDS Kerncorpus is the linguistic search engine DDC (Dialing DWDS Concordancer) (Sokirko 2003). DDC is a search engine developed specially to meet the needs of linguistic queries. It extracts metadata from the XML/TEI header and indexes the text. The input format is the common one-line per token format ("vertical file") where each annotation level such as lemma or part-of-speech is stored in a tab-separated column. The linguistic annotation together with the extraction of sentence boundaries is provided by TAGH morphology and the Moot tagger. The document delimiter corresponds to the sentence boundaries, thus Boolean queries operate on the sentence level and not on the entire document, as in commercial search engines. Input queries for DDC may consist of sequences of word forms, lexical categories, lemmas, thesaurus elements, or combinations of all four. Also supported are right and left-truncated searches, Boolean AND, OR, NOT searches, and interval searches (NEAR, FOLLOWED_BY), and regular expressions (except for negation). DDC allows also for filtering and sorting of metadata. Thus, it is possible to restrict searches to a specific text, author or text type and to sort the database by date or by sentence length.

The DWDS Kerncorpus can be queried for collocations, i.e. for sequences of words which co-occur more often than would be expected by chance such as *fried* and *potatoes*, *cutting* and *edge* or *warm* and *coat*. The statistic module operates on the complete DWDS Kerncorpus and computes collocations according to common statistic association measures including mutual information, t-score and log-likelihood.

Copyrights impose restrictions on the use of the Kerncorpus. First, texts with copyright restrictions are password-protected. They are still accessible at no charge but users must open an account and agree to use the corpus for non-commercial purposes only. Depending on the particular copyright agreement, smaller or wider contexts are displayed by the search engine (between 7 words and the paragraph), named entities are anonymized (e.g. for letters) or only samples are provided. For example, the Saur Verlag has granted the use of Karl Kraus' *Die Fackel,* but only 5% of the entire text can be shown via the web interface.
Second, most of the publishing houses ask for a guarantee that a complete text by a given author cannot be extracted. Currently, this is done by limiting all queries to 500 hits. We plan to admit larger result sets of up to 5000 hits in the future for general queries. However, queries with special author or title filters will be restricted to smaller result sets and the context for these texts will be reduced to one sentence. The first restriction avoids that the user extracts the entire book by formulating very general disjunctions, the latter avoids the concatenation of the entire text by queries iterating over sentences.
Third, 29% of the Kerncorpus (measured in tokens) can only be accessed internally. Even though the texts themselves are invisible from the public query interface, they are present in the corpus and available for statistic queries.

12. Conclusion and further work

The DWDS Kerncorpus is the first reference corpus for the German language of the 20th century. It is balanced, equally distributed over all periods of the 20th century, and comparable in size to the British National Corpus. The corpus is lemmatized and part-of-speech tagged, enabling linguistic queries. Furthermore, several standard tools for the computation of collocations are integrated into the web based query interface. A more fine-grained linguistic filtering is planned: we plan to annotate the corpus by means of a shallow parser. This will enable users to formulate phrase or chunk-based searches. We further intend to build a tree bank from a subset of the DWDS Kerncorpus.

The DWDS Kerncorpus is used as a source for language-based work not only by linguists but also by historians and academics from a range of disciplines, by translators, and by many lay people for various purposes. Moreover, the Kerncorpus can be used as a resource for psychological and psycholinguistic research. In a recent study, we demonstrated with case-sensitive frequency norms computed from the DWDS corpus that fixation durations during reading of sentences reveal theoretically predicted effects of influences of the frequency and the informativeness of the initial letters of the upcoming, not yet fixated word (Geyken, Hanneforth, & Kliegl, in prep., Kliegl, Geyken, & Hanneforth, in prep.). These effects were not reliable when norms were computed from the unbalanced and much smaller CELEX corpus (Baayen, Piepenbrock, & Rijn, 1993). DWDS and CELEX norms did not differ in the prediction of fixation durations of fixated word properties, although, as expected from corpus sizes, DWDS-based visualizations were considerably less noisy than CELEX-based ones. The DWDS norms will be of great use for the further expansion of a reading-based eye-movement corpus (Kliegl, Nuthmann, & Engbert, 2006) and the further development of a computational model of eye-movement control during reading (Engbert, Nuthmann, Richter, & Kliegl, 2005).

The DWDS Kerncorpus can be accessed at no costs from the project website. Currently, more than 6,000 users are registered for the use of the copyrighted part of the DWDS Kerncorpus and many others use the freely available part of the corpus. The number of requests received and searches carried out have already surpassed all expectations. However, the level of interest expressed makes it obvious that the corpus needs constant updating and expansion. We intend to gradually add more texts from the 20th century and to extend the corpora to the 21st century and to texts before 1900. In connection with these plans, we expect to intensify cooperation with publishing houses and with other research projects. As an example, we cite the recent cooperation with Compact Memory, a project that digitizes Jewish periodicals of the 19th and 20th centuries.

Recent studies have shown that a 100 million word corpus is not sufficient for exploring phenomena such as the use of rare words and rare combinations of words (Kilgarriff and Grefenstette 2003, Geyken 2004), and that larger data collections, even though they may be noisy, provide better results than the ones based on estimates from smaller and cleaner corpora (e.g. Keller and Lapata 2003). For this reason we intend to enlarge our opportunistic corpus. We are currently negotiating with several supra-regional newspapers and magazines about the restricted use of parts of their electronic archives.
Apart from the expansion of the corpus, future work will draw on the comparison of the DWDS Kerncorpus and the recently created balanced web-based corpora (e.g. Sharoff 2004). On the one hand, such corpora are quite different: the DWDS Kerncorpus is a reference corpus and covers a time interval of 100 years whereas internet corpora are monitor corpora. The DWDS Kerncorpus has a stable statistical basis whereas the URLs of

the internet corpus are subject to constant change. And third, while copyright restrictions make it a challenge to keep the DWDS Kerncorpus up to date, this is comparatively easy for an internet corpus. On the other hand, both corpora are balanced corpora of the language. Despite all the aforementioned differences, it will be interesting to compare both corpora on both a lexical and morphological as well as a syntactic level.

Acknowledgements

I am grateful to Ines Rehbein for her help with the implementation of the sampling procedure and to Patrick Hanks for discussion and comments on an earlier draft of this paper.

Bibliography

Bergenholtz, Henning and Joachim Mugdan (1990): 'Formen und Probleme der Datenerhebung II: Gegenwartsbezogene synchronische Wörterbücher'. In: Hausmann, Franz Josef/ Reichmann, Oskar/ Wiegand, Herbert Ernst/ Zgusta, Ladislav (eds.): *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, 2. Teilband. Berlin/New York: de Gruyter [Handbücher zur Sprach- und Kommunikationswissenschaft; 5,2], pp. 1611–1625.

Biber, Douglas (1994): 'Representativeness in corpus design', in: Zampolli, Antonio/Calzolari, Nicoletta/Palmer, Martha (eds.): *Current Issues in Computational Linguistics: In Honour of Don Walker* (=Linguistica Computazionale IX-X). Pisa: Giardini/Dordrecht: Kluwer, pp. 377–407.

Engbert, R., Nuthmann, A., Richter, E.M., & Kliegl, R. (2005). 'SWIFT: A dynamical model of saccade generation during reading'. *Psychological Review*, 112, 777-813.

Fellbaum, Christiane (2004): 'Idiome in einem Digitalen Lexikalischen System'. In: *Zeitschrift für Literaturwissenschaft und Linguistik*. 34. Jg., Heft 136, 56–7.

Geyken, Alexander, Alexey Sokirko, Ines Rehbein and Christiane Fellbaum (2004): *What is the optimal corpus size for the study of idioms*? DGfS-Jahrestagung. Mainz (D), 25.–27.2. 2004.

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

Geyken, A., Hanneforth, T., & Kliegl, R. (in prep). 'Corpus matters: A comparison of DWDS and CELEX lexical and sublexical frequency norms for the prediction of fixation durations during reading'.

Hanks, Patrick (2005): 'Johnson and modern lexicography'. In: *International Journal of Lexicography*. 2005 18(2):243–267. Oxford. University Press.

Harm, Volker (2005): 'Perspektiven auf die sprachhistorische Lexikographie nach dem deutschen Wörterbuch'. In: *Zeitschrift für germanistische Linguistik* (ZGL) 33, 2005, 92–105.

Heid, Ulrich , Worsch, Wolfgang, Evert, Wermke, Dougherty, Vincent (2000): 'Computational linguistic tools for semi-automatic corpus-based updating of dictionaries'.

In: *Proceedings of Second International Conference on Language Resources and Evaluation (LREC 2000)*.

Heid, Ulrich , Bettina Säuberlich, Esther Debus-Gregor, Werner Scholze-Stubenrecht (2004): 'Tools for upgrading printed dictionaries by means of corpus-based lexical acquisition'. In: *Proceedings of LREC 2004*, Lisbon (Portugal), pp. 911–914.

Jurish, Bryan (2003): *A Hybrid Approach to Part-of-Speech Tagging*, Final report, Project Kollokationen im Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin.

Keller, Frank and Mirella Lapata (2003): 'Using the Web to obtain frequencies for unseen bigrams'. In: *Computational Linguistics* 29:3, 459–484.

Kilgarriff, Adam and David Tugwell (2004): 'Sketching words'. In: *Proceedings of Euralex*, Lorient, France, July 2004, pp. 105–116.

Kilgarriff, Adam and Gregory Grefenstette (2003). *Introduction to the special issue on the web as corpus*. Computational Linguistics 29:3, 333-348.

Klein, Wolfgang (2004): 'Das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS)'. In: J. Scharnhorst (ed.), *Sprachkultur und Lexikographie*. Berlin, Peter Lang, pp. 281–311.

Klein, Wolfgang and Alexander Geyken (2000): 'Projekt "Digitales Wörterbuch der deutschen Sprache des 20. Jh."'. In: *Jahrbuch der BBAW 1999*, Berlin, Akademie Verlag, pp. 277–289.

Kliegl, R., Geyken, A., & Hanneforth, T. (in prep). 'Parafoveal effects of word frequency, familiarity, and regularity on fixation durations in reading.'

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). 'Tracking the mind during reading: The influence of past, present, and future words on fixation durations'. *Journal of Experimental Psychology: General*, 135.

Kučera, H. und W. N. Francis (1967): *Computational Analysis of Present-day English*. Brown University Press. Providence, Rhode Island.

Schmidt, Hartmut (1997): 'Plädoyer für eine moderne korpusbasierte deutsche Wortschatzforschung'. In: *Zeitschrift für Literaturwissenschaft und Linguistik*. 27. Jg. (1997), Heft 106, 19–29.

Schulz Matthias (2001): 'Einzelwortbeschreibung und Wortschatzbeschreibung'. In *Sprachwissenschaft* 26, 41–58.

Sharoff, Serge (2004): 'Towards basic categories for describing properties of texts in a corpus'. In: *Proceedings of LREC 2004*, Lisbon, Portugal, May 2004, 1743–1746.

Sinclair, John (1996): 'EAGLES Preliminary recommendations on Corpus Typology'. EAG-TCWG-CTYP/P. Version of May 1996.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants and Hans Uszkoreit, 1997: 'An annotation scheme for free word order languages. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. Washington, DC, USA.

Sokirko, A. (2003): 'DDC – A search engine for linguistically annotated corpora'. In: *Proceedings of Dialogue 2003*, Protvino, Russia, June 2003.

Telljohann, Heike, Erhard W. Hinrichs and Sandra Kübler (2004): 'The TüBa-D/Z Treebank — Annotating German with a context-free backbone. In: *Proceedings of LREC 2004*. Lisbon, Portugal.

Appendix: Distribution of tokens per decade and text genre (version 0.95)

Text type	Decade (1900–2000)	tokens
literature	1	2542807
literature	2	3259726
literature	3	3212220
literature	4	3261446
literature	5	2078014
literature	6	3234186
literature	7	2227026
literature	8	2391338
literature	9	2023919
literature	10	2332459
other fiction	1	2286829
other fiction	2	2421260
other fiction	3	2438427
other fiction	4	2151094
other fiction	5	2442199
other fiction	6	2402293
other fiction	7	1292670
other fiction	8	1363232
other fiction	9	2419838
other fiction	10	2414668
science	1	2374435
science	2	2725471
science	3	2444098
science	4	2550222
science	5	2147141
science	6	2596157
science	7	2390911
science	8	2371863
science	9	2399778
science	10	2371571
newspaper	1	2346961
newspaper	2	2546206
newspaper	3	3139119
newspaper	4	3138888
newspaper	5	2842066
newspaper	6	2804838
newspaper	7	2803179
newspaper	8	2802643
newspaper	9	2804187
newspaper	10	2805608

[1] The DWDS Kerncorpus was funded by the Deutsche Forschungsgemeinschaft (DFG) between 2000 and 2003 as well as by the Berlin-Brandenburgische Akademie der Wissenschaft.