The Database for Spoken German – DGD2

Thomas Schmidt
Institut für Deutsche Sprache
R5, 6-13, D-68161 Mannheim
E-mail: thomas.schmidt@ids-mannheim.de

Abstract

The Database for Spoken German (Datenbank für Gesprochenes Deutsch, DGD2, http://dgd.ids-mannheim.de) is the central platform for publishing and disseminating spoken language corpora from the Archive of Spoken German (Archiv für Gesprochenes Deutsch, AGD, http://agd.ids-mannheim.de) at the Institute for the German Language in Mannheim. The corpora contained in the DGD2 come from a variety of sources, some of them in-house projects, some of them external projects. Most of the corpora were originally intended either for research into the (dialectal) variation of German or for studies in conversation analysis and related fields. The AGD has taken over the task of permanently archiving these resources and making them available for reuse to the research community. To date, the DGD2 offers access to 19 different corpora, totalling around 9000 speech events, 2500 hours of audio recordings or 8 million transcribed words. This paper gives an overview of the data made available via the DGD2, of the technical basis for its implementation, and of the most important functionalities it offers. The paper concludes with information about the users of the database and future plans for its development.
Keywords: speech database, corpus platform, speech corpora

1. Introduction

The Database for Spoken German (Datenbank für Gesprochenes Deutsch, DGD) is the central platform for publishing and disseminating spoken language corpora from the Archive of Spoken German (Archiv für Gesprochenes Deutsch, AGD, Stift & Schmidt 2014) at the Institute for the German Language in Mannheim. Its first version (Fiehler & Wagener 2005) has been available since 2003. The current version (DGD2, http://dgd.ids-mannheim.de), described in this paper and first published in December 2012, is a complete redevelopment. It puts the database on a new technical basis, substantially extends the functionalities for browsing, querying and downloading data, and integrates the Research and Teaching Corpus of Spoken German (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK) as a new large resource of spoken German.

This paper gives an overview of the data made available via the DGD2 (section 2), of the technical basis for its implementation (section 3), and of the most important functionalities it offers (section 4). The paper concludes with information about the users of the database (section 5) and future plans for its development (section 6).

2. Corpora

The corpora contained in the DGD2 come from a variety of sources, some of them in-house projects, some of them external projects. Most of the corpora were originally intended either for research into the (dialectal) variation of German or for studies in conversation analysis and related fields. The AGD has taken over the task of permanently archiving these resources and making them available for reuse to the research community.[1]

To date, the DGD2 offers access to 19 different corpora, totaling around 9000 speech events, 2500 hours of audio recordings or 8 million transcribed words.

Among these corpora are, on the one hand, several larger corpora documenting dialects and other regional variation in German. Most importantly, these include the corpus "Deutsche Mundarten" (aka "Zwirner-Korpus"), assembled in the 1950s and 1960s and containing 5795 recordings (around 1000 hours) of semi-structured interviews with dialect speakers evenly distributed over the area of the former Federal Republic of (Western) Germany, two satellite corpora following the same design as the Zwirner-Korpus, but covering the formerly German territories in Eastern Europe (corpus "Ehemalige Deutsche Ostgebiete", 981 recordings, around 500 hours, Bellman & Göschel 1970) and the former Democratic Republic of (Eastern) Germany[2] ("DDR-Korpus", 1625 recordings, around 400 hours, Schädlich & Eras 1965), respectively, and the corpus "Deutsche Umgangssprache" (aka "Pfeffer-Korpus", Pfeffer & Lohnes 1984) documenting regional variation in non-dialectal language (398 recordings, around 80 hours).

On the other hand, the DGD2 contains a number of conversation corpora documenting spontaneous verbal interaction in different discourse domains, such as the "Freiburger Korpus" (222 recordings, around 70 hours, Engel & Vogel 1975) and the corpus "Dialogstrukturen" (70 recordings, around 15 hours, Berens et al. 1976), both from the 1970s.

[1] For legal as well as technical reasons, not all resources hosted by the AGD are suitable for publication via the DGD2. However, the resources published in the DGD2 make up by far the larger part of the overall AGD resources.

[2] This corpus is currently being processed for integration in the DGD2. The first part is expected to be ready for publication in the course of 2014.

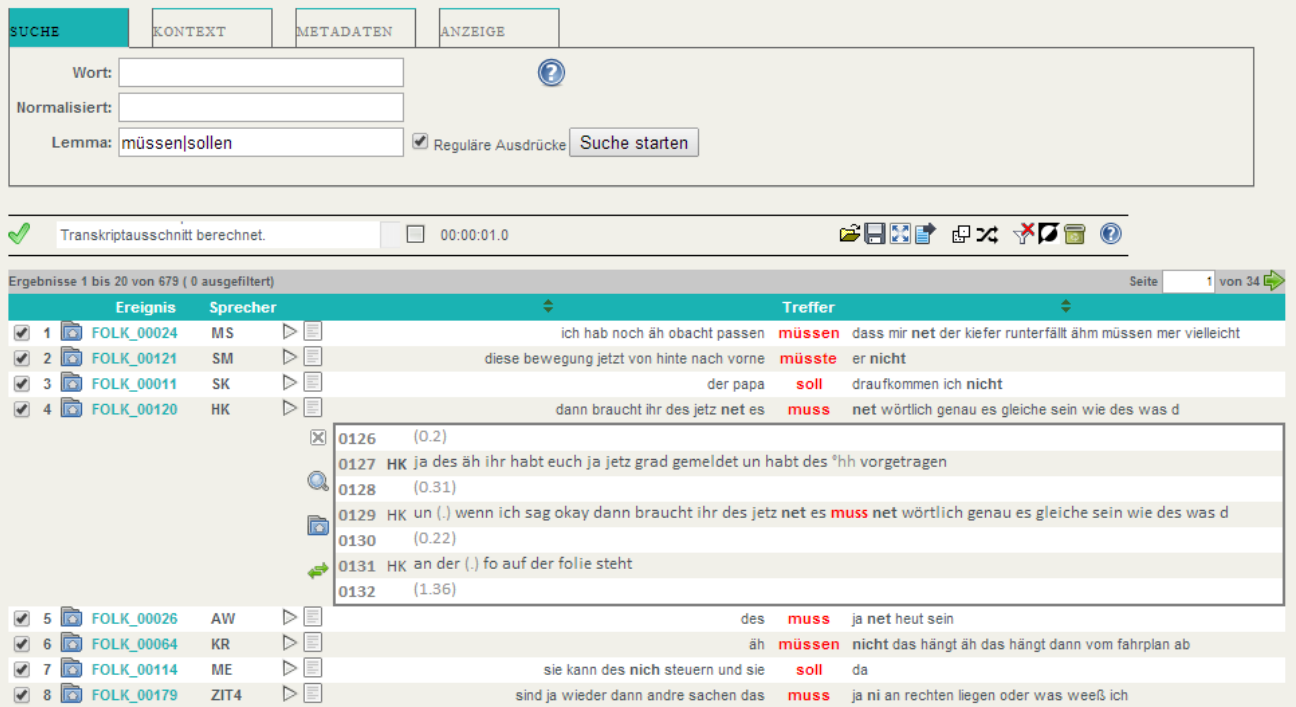

Figure 1: Query for lemmas *müssen*/*sollen* with *nicht* in context, resulting KWIC Concordance with transcript excerpt

More recently, the AGD has started to build up the Research and Teaching corpus of Spoken German (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK, Deppermann & Hartung 2012, Schmidt 2014), a large German conversation corpus currently comprising around 120 recordings (around 100 hours) of spontaneous conversation from different private, institutional and public settings. FOLK is also available to the public via the DGD2.

For a complete overview of the corpora in the DGD2, see appendix.

In line with the institute's mission to become a central provider for German corpus data, the DGD2 must also be ready to integrate new data from external projects. Currently, several such external corpora are being curated, the legacy of Michael Clyne's work on German in Australia (Clyne 1981) among them.

3. Technical basis

The DGD2 has dedicated data models for documenting corpus, speech event and speaker metadata (Gasch et al. 2008) and a dedicated data model for representing transcription data. The latter is also the basis for the FOLKER transcription editor (Schmidt 2012) and is compatible with the most important other data models currently used for multimedia annotation (such as ELAN, EXMARaLDA, Praat, Transcriber), and also with the TEI guidelines (Schmidt 2011). Both data models are represented in XML files during corpus creation.

For corpus dissemination, the DGD2 uses an object-relational (Oracle) database to incorporate and index these XML files on the basis of the corresponding XML schemas. This approach combines advantages of relational database technology (fast access, controlled storage) while keeping it easy to maintain and update corpora created as bundles of XML files. Only for the most performance critical part of the application, the query for individual tokens, a relational table is used. For atomic queries to the database, a low-level API written in PL/SQL is provided. This API is then called via http by a Java EE web application (running in a TomCat environment) which combines atomic queries into more complex ones and takes care of communication with the user web interface. This "middle tier" is also suitable for delivering web services to other applications (e.g. a federated search in CLARIN).

Data for the FOLK corpus are currently created with FOLKER as transcripts in modified orthography according to the GAT conventions (Selting et al. 2009). Added to this transcription level are a standard orthographic normalization, and a lemmatization and POS tagging using TreeTagger (Schmid 1995). Transcripts in FOLK are manually aligned with the recording with a granularity of three to five seconds during transcription.

Transcription systems, and also the portion of the audio which has been transcribed at all, vary for the other corpora in the DGD2. For example, most variation corpora are transcribed in standard orthography, recordings from the Pfeffer-Korpus have been completely transcribed, but there are transcriptions for only about 40% of all recordings in the Zwirner-Korpus. Where transcripts exist, most of them were aligned automatically on the word level. Currently, the WebMAUS tool (Kisler/Schiel/Sloetjes 2012) is used for that purpose.

4. Database functionality

The DGD2 addresses a diverse audience ranging from phoneticians over dialectologists, conversation analysts to corpus and computational linguists and speech technologists. In order to provide all these user groups

with an adequate access to the data, the DGD2 offers three principal modes of access which are described in the following subsections.

4.1. Online browsing

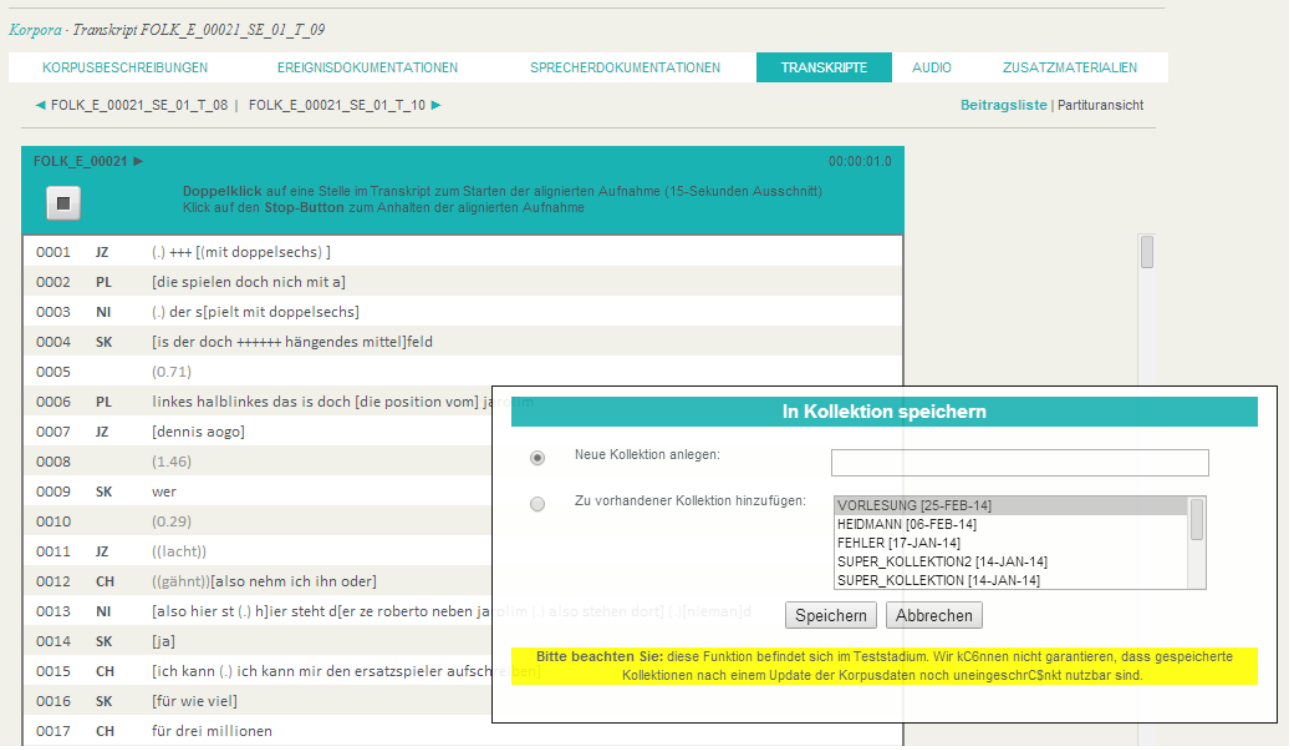

Figure 2: Browsing a transcript and saving an excerpt in a collection

Online browsing allows users to read through metadata and transcriptions and to listen to audio recordings via the web interface of the database. The different data types – metadata, recordings, transcripts and additional material – are always linked with one another so that it is possible to directly navigate from a documentation to the corresponding transcript(s) or recordings, to playback recordings in sync while reading a transcript and to call up information about speakers from each occurrence in a transcript. For FOLK, transcripts can be visualized not only in a line-for-line layout, but also in a musical score layout which makes it easier to study overlapping speaker contributions.

Online browsing is both a way of getting acquainted with the resources, i.e. of judging their content and quality, and a way of carrying out qualitative (transcript) studies on the data, which is the typical way for conversation analysts to work with spoken language data.

In order to support the "manual" gathering of example excerpts, the DGD2 offers a collection functionality (see figure 2). A click on any place in a transcript will add the corresponding excerpt to a collection which users can permanently save and retrieve in the database.

4.2. Online querying

Online querying allows users to systematically query communication and speaker metadata as well as the transcription with its different annotation levels. Currently, there are three different querying modes.

Full text queries (see figure 3) are the simplest query option. With the help of a query language which comprises conjunction, disjunction, distance and stemming operators and fuzzy search, users can search for text strings in the full text of metadata documentations and transcripts regardless of their internal structure. Full text queries are suitable for coarse searches on the document level (for instance, for finding all transcripts which contain certain words). Because they do not make any assumptions about the internal structure of the data, full text queries also ensure a minimum level of searchability over the whole database, i.e. also over those transcripts which are only stored as unstructured text documents.

Figure 3: Full text query on transcripts

For more advanced purposes, two types of structured query are available which operate on the XML documents for metadata and transcripts, respectively.

Structured metadata queries (see figure 4) allow users to select from the whole database those datasets whose metadata meet certain search criteria – for example: only speech events of type 'interview' recorded in Northern Germany, only speech events with female speakers younger than 40 years with a university degree, or any combination of such speech event and speaker properties. The result of a structured metadata query can be manually

inspected (enabling users to sort out false positives, for example) and saved permanently inside the database. When used in conjunction with a structured token query (see below), the result of a metadata query can be used as a 'virtual corpus', i.e. a preselection of datasets.

Figure 4: Structured metadata query

Structured token queries, finally, are the most powerful search mechanism so far implemented in the DGD2. A structured token query is initiated by specifying properties of individual tokens, presently including the transcribed form, the orthographically normalized form, and the lemmatized form. These properties can be formulated as regular expressions and combined at will. The same mechanism can be used to specify properties of tokens in the left or right context of matching tokens. Additionally, metadata properties for speech events and speakers can be specified in order to further restrict the query space (the same effect can be achieved by using a virtual corpus, see above).
Results of structured token queries are displayed as KWIC concordances in which matching tokens and matching context are highlighted (see figure 1). For each line of the concordance, links are given to a) display metadata about the speech event, b) display metadata about the speaker, c) playback the corresponding part of the aligned audio, d) download the excerpt on the local computer (see below) and e) show the matching item in the context of the full transcript. Metadata properties for each result can be displayed in additional columns of the KWIC table. Since queries on spoken language data are often over-selective, individual lines of the concordance representing, for instance, false positives can be manually deselected and removed. Further options for interacting with the KWIC concordance and post-processing search results include:

- sorting the concordance according to any of its columns,
- extracting a random sample from a larger search result,
- randomly shuffling a search result,
- inverting and resetting filters,
- saving search results permanently in the database,
- exporting search results in a form suitable for further processing, e.g. in spreadsheet applications, or for integration into printable documents.

Structured token queries support a corpus linguistic approach to the DGD2 data. They include most functionality typically present in corpus platforms for written language and supplement this with functionality specific to spoken data.

4.3. Download

The functionality for downloading (excerpts of) transcripts and the corresponding recordings and metadata allows users to process individual datasets on their local computers. Transcripts can be opened with the FOLKER or OrthoNormal tool (see figure 5) and be exported to most other common formats (e.g. EXMARaLDA, Praat, ELAN) typically used for working with spoken language data.

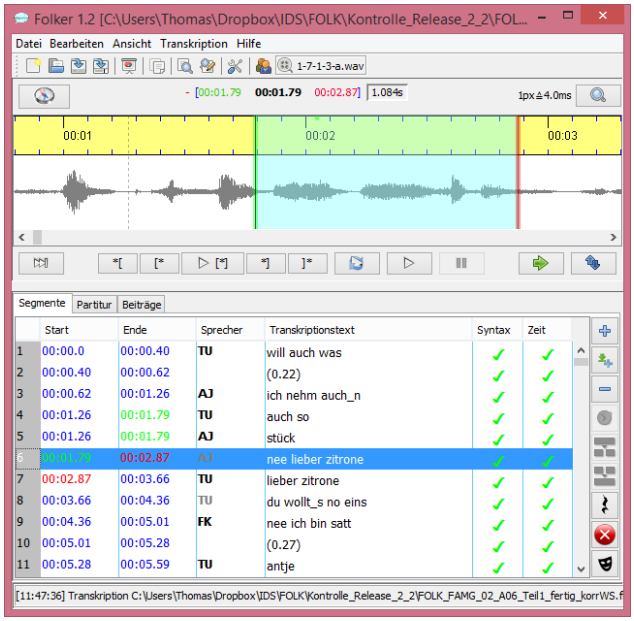

Figure 5: Transcript opened in FOLKER

This approach is best suited wherever users want to investigate the data for phenomena that are not represented in the available annotations (e.g. prosodic features). The download functionality can be called up for individual entries of a collection or a KWIC concordance (see above), but also for selected entire datasets of the corpora FOLK, PF and ZW.

5. Users

In line with the mission of the Institute, usage of the DGD2 is free to members of academia for non-commercial research and teaching purposes. A one-time registration is required in order to ensure that users are informed about the terms of use. At the time of writing, the DGD2 has approximately 2000 registered users, and the number of users is constantly growing. Considering that the database has been online for less than two years, this clearly demonstrates that there is a great interest in the linguistic communities involved for spoken language corpora of the type offered in the DGD2.
A preliminary analysis of registrations reveals that a substantial proportion of them (more than 20%) come from non-German speaking countries where researchers use the data, among others, in the teaching of German as a foreign language. With more than 60% of all requests to the Tomcat server, FOLK is by far the most popular corpus in the database, followed by the large variation corpora (ZW, PF, about 20% of all requests) and the "historical" conversation corpora (FR and DS, about 10% of all requests).
A more detailed user analysis, involving a systematic survey and detailed case studies, is in the planning stage.

This user analysis will also help us to determine priorities for future development of the DGD2.

6. Outlook

Being located at the Institute for the German Language, the DGD2 has a good perspective of establishing itself as a permanent instrument for making accessible corpora of spoken German. The Archive of Spoken German is constantly acquiring new data from external projects, most of which (presupposing legal authorization of the data) can and will be published via the DGD2.

Construction of the FOLK corpus is also ongoing, with the latest release (March 2014) attaining the 100h (or 1.000.000 transcribed tokens) mark. The corpus is then meant to grow by at least 20 additional hours per year.

The functionality of the database is also continually extended. Current work focuses on more advanced search functionality, where we are considering an integration of query languages widely used in corpus linguistic communities (like CQP or Poliqarp), and the extension of the existing functionality for personal user spaces which allows users to store and share analysis results.

7. Acknowledgements

The DGD2 is developed by members of the Archive of Spoken German (AGD) at the Institute for the German Language (IDS) in Mannheim. Parts of this paper are based on Schmidt, Dickgießer & Gasch (2013).

8. References

Bellmann, G. and Göschel, J. (1970). *Tonbandaufnahmen ostdeutscher Mundarten 1962-1965*. Gesamtkatalog. Marburg (= DDG 73).

Berens, F.-J.; Jäger, K.-H.; Schank, G. and Schwitalla, J. (1976). *Projekt Dialogstrukturen. Ein Arbeitsbericht.* Heutiges Deutsch I/12. München: Hueber

Clyne, M. (1981). *Deutsch als Muttersprache in Australien*. Wiesbaden: Franz Steiner Verlag.

Engel, U. and Vogel, I. (1975) (Eds.). *Gesprochene Sprache*. Bericht der Forschungsstelle Freiburg. Tübingen: Narr.

Gasch, J.; Brinckmann, C. and Dickgießer, S. (2008). *memasysco: XML schema based metadata management system for speech corpora*. In: Proceedings of LREC 2008, Marrakech, Morocco.

Kisler, T.; Schiel, F. and Sloetjes, H. (2012). *Signal processing via web services: the use case WebMAUS*. In: Proceedings Digital Humanities 2012, Hamburg, Germany, Hamburg, pp. 30-34.

Pfeffer, J.; Lohnes, W. (eds.)(1984). *Grunddeutsch. Texte zur gesprochenen deutschen Gegenwartssprache*. (Phonai, Bde. 29 u. 30) Tübingen: Niemeyer

Schädlich, H.-J. and Eras, H. (1965). *Bericht über die Tonbandaufnahmen der deutschen Mundarten in der Deutschen Demokratischen Republik.* In: Berichte über dialektologische Forschungen in der Deutschen Demokratischen Republik. Berlin. S. 24-27

Schmid, H. (1995). *Improvements in Part-of-Speech Tagging with an Application to German.* Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schmidt, T. (2011). *A TEI-based Approach to Standardising Spoken Language Transcription.* In: Journal of the Text Encoding Initiative (1).

Schmidt, T. (2012). *EXMARaLDA and the FOLK tools.* In: Proceedings of LREC 2012, Istanbul, Turkey.

Schmidt, T. (2014). *The Research and Teaching Corpus of Spoken German – FOLK*. In: Proceedings of LREC 2014, Reykjavik, Iceland.

Schmidt, T.; Dickgießer, S. and Gasch, J. (2013). *Die Datenbank für Gesprochenes Deutsch - DGD2.* Mannheim: Institut für Deutsche Sprache. URN: urn:nbn:de:bsz:mh39-12747.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertz-lufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., Uhmann, S. (2009). *Gesprächsanalytisches Transkriptionssystem 2 (GAT 2).* In: Gesprächsforschung (10), pp. 353 - 402.

Appendix: Overview of DGD2 corpora

Conversation corpora			
BR	Biographische und Reiseerzählungen	Narrative interviews with young speakers of former East Germany	7 speech events 7 audio recordings (5:30h) 7 transcripts (11443 tokens) 24 documented speakers
DS	Korpus "Dialogstrukturen"	Different types of spontaneous interaction between speakers of standard German	70 speech events 70 audio recordings (15:18h) 70 transcripts (142661 tokens)
EK	Elizitierte Konfliktgespräche	Elicited conflict interaction between mothers and their teenage daughters	107 speech events 138 audio recordings (12:23h) 138 transcripts (162123 tokens)
FOLK	Forschungs- und Lehrkorpus Gesprochenes Deutsch	Different types of spontaneous interaction in private, institutional and public settings.	137 speech events 149 audio recordings (101:12h) 263 transcripts (965652 tokens) 360 documented speakers
FR	Grundstrukturen: Freiburger Korpus	Different types of spontaneous interaction between speakers of standard German	222 speech events 222 audio recordings (68:06h) 221 transcripts (593196 tokens)
SA	Kindersprache: Saarbrücker Korpus	Child-Adult interaction	48 speech events 48 audio recordings (04:33 h)
Corpora documenting variation of German within Germany (and Austria and Switzerland)			
BB	Deutsche Mundarten: Kreis Böblingen	Narrative interviews with dialect speakers from the area of Böblingen (SW Germany)	73 speech events, 73 audio recordings (42:28h)
HL	Deutsche Hochlautung	Excerpts from news and other broadcasts and from press conferences	27 speech events 9 documented speakers 37 audio recordings (01:57h) 27 transcripts (9744 tokens)
KN	Deutsche Standardsprache: König-Korpus	Reading texts: Excerpts from the German Grundgesetz	43 speech events 43 Sprecher 37 audio recordings (05:48h) 43 transcripts (41573 tokens)
PF	Deutsche Umgangssprachen: Pfeffer-Korpus	Narrative interviews with speakers from different areas of former West and East Germany, Austria and Switzerland	398 speech events 403 documented speakers 398 audio recordings (79:15h) 398 transcripts (646492 tokens)
OS	Deutsche Mundarten: ehemalige deutsche Ostgebiete	Narrative interviews with dialect speakers from former German territories in Eastern Europe	981 speech events 989 documented speakers 981 audio recordings (462:05h) 280 transcripts (833159 tokens)
SV	Deutsche Mundarten: Südwestdeutschland und Vorarlberg	Narrative interviews with dialect speakers from the Vorarlberg and Liechtenstein region (Austria)	242 speech events 242 documented speakers 242 audio recordings (72:06h)
SW	Deutsche Mundarten: Schwarzwald	Narrative interviews with dialect speakers from the Schwarzwald area (SW Germany)	126 speech events 122 documented speakers 126 audio recordings (37:31h)
ZW	Deutsche Mundarten: Zwirner-Korpus	Narrative interviews with dialect speakers from different (mostly rural) areas of former West Germany.	5795 speech events 5887 documented speakers 5795 audio recordings (1076:56h) 2311 transcripts (3754039 tokens)
Corpora documenting varieties of German outside Germany and Austria			
IS	Emigrantendeutsch in Israel	Interviews with Jewish speakers of German which emigrated from Germany to Israel in the 1930s.	142 speech events 165 documented speakers 142 audio recordings (231:04 h) 20 transcripts (309739 tokens)
ISW	Emigrantendeutsch in Israel: Wiener in Jerusalem	Interviews with Jewish speakers of German which emigrated from Wien, Austria, to Israel in the 1930s.	25 speech events 21 documented speakers 25 audio recordings (43:37h) 24 transcripts (285664 tokens)
ISZ	Zweite Generation deutschsprachiger Migranten in	Interviews with children of the German speaking emigrants from the corpora IS and ISZ	60 speech events 57 documented speakers

	Israel		60 audio recordings (109:00h)
MV	Binnen- und auslandsdeutsche Mundarten: Varia	Various interviews with speakers of different varieties of German	183 speech events 184 documented speakers 183 audio recordings (09:25h)
SR	Slawische Mundarten im Ruhrgebiet	Narrations of Polish, Slowenian and Ukrainian immigrants in the Ruhrgebiet area	23 speech events 23 documented speakers 23 audio recordings (06:40 h)