Originally published in: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002). – Las Palmas: University of Las Palmas, 2002. Pp. 1-4.

Third International Conference on Language Resources and Evaluation
May 29-31, 2002

International Workshop on Resources in Tools and Field Linguistics
May 26 and 27, 2002

Las Palmas, Canary Islands - Spain

ONLINE ACCESS TOOLS FOR SPOKEN GERMAN: THE RESOURCES OF THE DEUTSCHES SPRACHARCHIV IN A DATABASE

Peter Wagener
Institut für Deutsche Sprache

Abstract

This paper shows some details of the modernization of the Deutsches Spracharchiv (DSAv). It explores some future possibilities of linguistical documentation and analysis using the Web. The Institut für Deutsche Sprache (IDS) in Mannheim is the central institution for linguistic research in Germany. The DSAv in the IDS is the center for documentation and research of spoken German. These archives include the largest collection of sound recordings of spoken German (dialects and colloquial speech, including e.g. lots of extinct dialects of former German territories in Eastern Europe) - altogether more than 15,000 sound recordings. The lacking clarification and accessibility of this data material has been felt as an essential deficit. The opportunity to edit the sound signal digitally offers a much easier access to spoken language. Through the integration of the already existing information about the corpora and the transcribed texts in an information- and full text databank, as well as the linking of the data with the acoustic signal (alignment), arises a data-pool with considerably better documentation of the materials and a fast direct grasp of the recorded sounds. Thus, the DSAv initiates totally new research questions for the work at the IDS, as well as for linguistics altogether.

0. THE PROJECT "DATABASE OF SPOKEN GERMAN"

In summer of 1997, the Deutsches Spracharchiv (DSAv, German Archive of Spoken Language) was able to begin the main step of its modernization. As a part of the program "Foundation and development of archives", sponsored by the Volkswagen Research Fund, the "Computerbased registration and organization of the sound recordings concerning spoken language from the DSAv" was enabled. The database "Spoken German" has been installed. The goal of this project was to base the oganization and archivation of the material on electronic devices in order to extend their accessibility for scientific research all over the world. With the development of the database "Spoken German", this goal has been achieved. It is now available on the internet, allowing worldwide research among the material of the DSAv.

(http://www.IDS-Mannheim.de/DSAv)

The database includes documentary data about the corpora and the recordings, the sound recordings, as far as they have been digitized, and transcripts of the recordings, as ar as they exist.

A search among the documentary data allows a general overview over the material and traces the interactions corresponding to the criteria of the search. If transcripts of these interactions are available, they can be shown on the screen. The transcripts have been aligned with the corresponding sound recordings on the level of words, so that navigation within the recordings and efficient research is possible. This way, the search leads to extracted verifications for any phenomenon examined.

Organizing the material within a database enables the user to characterize the recordings by increasingly differentiated traits. Due to the large number of search options, access to the information about the material also becomes many times faster as well as more variable, exact and extended. Certain traits or types of material can be searched for aimfully. The cataloguing, an anachronistic defile for the use of the material, can thus now be replaced by more efficient procedures.

1. THE ONLINE PRESENTATION OF THE DATABASE "SPOKEN GERMAN"

Starting out on a general homepage, the user can choose from a number of links to gain information on the build-up of the database, the DSAv and its materials, its history and the staff, but especially on the numerous opportunities of research based on the sound recordings and transcripts accessible so far.
In order to achieve the goals of the project, the following tasks have been worked on in detail:

- Preparation of the documentary data, so that each interaction can be describe by a defined sentence of information
- Preparation of the transcripts in stock at the archive – partly machine- or even handwritten – by digitizing them and converting them into a formate of archivation fit for computer processing. Then, different versions of the transcripts can be created automatically according to their various applications, such as the fulltext database.
- Preparing those sound recordings already available in digital form for transformation into a formate defined for the database "Spoken German"
- Preparation of digital transcripts and sound recordings for the half-automatic method of synchronisation developed by the IDS (alignment)
- Development of a homogenous system to prepare of the different archive material corresponding to the necessities of a database
- Developing a database

Three different versions of the database "Spoken German" are accessible: an internal one, a public one and one for external scientific use. All three versions can so far gain access to and search within about 9000 documentations of interactions The public version aditionally contains thirty exemplary digitized sound recordings and the corresponding transcripts. Here, the recordings and transcripts are aligned by words. The version for external use so far comprises some 4000 digitized sound recordings and 3000 transcripts, some 2000 of which have been synchronized to match the corresponding recordings so far.

2. THE TECHNICAL CONCEPTION OF THE DATABASE "SPOKEN GERMAN"

The DSAv has the assignment to archivate the records of interactions, mainly sound recordings and corresponding transcripts, to systematically register them and to set up ways of access to this material which allow its efficient use. These materials are parts of different corpora, originating from project with varying research questions and working methods.
The archivation, registration and provision of the heterogenous material of spoken language is technically managed by adjusting it into a form suitable for computer processing. Moreover, they are technically unified while their special contental qualities are being preserved. The archived material is divided into three different parts - documentations, transcripts and sound recordings - , which are being adjusted separately.

2.1. PREPARATION OF THE MATERIAL

In order to administrate the material of the various corpora, a new integrating system has been created. In the crux, new naming conventions for all materials to be administered in the archive have been developed. The signatures were defined to allow systematic computer processing using different operating systems. But they also simplify access to and research among the material for the user.

2.1.1. DOCUMENTATION

The documentation of interactions as wee as the material resulting from them are administrated by a relational database created as a part of the project, which can be handled using SQL. In this database, among other things, the status of processing of each transcript and each sound recording is recorded. While concepting the structure of the database, a special emphasis was laid on allowing the history of adaption of the archived material to be understood later. Therefore, the relations between predecessing and following states of processing are shown in the database.

2.1.2. TRANSCRIPTS

As a part of the organization of the archive, the materials from various corpora are unified wthout falsification of their special contental qualities.
The transcripts from different corpora administrated in the archive differ in technical aspects. In order to administrate and use them within an archive using identical means, a unified technical formate has been developed for the transcripts. It is so flexible, that even specific types of notation differing among the corpora can be preserved. The formate of archivation for transcripts was defined specifically for the DSAv by means of XML. The semantic used of annotations in the transcripts has to be documented only once per corpus. As a result of this proceeding, the DSAv is well prepared to integrate further corpora developed in other contexts without losing their special qualities.

2.1.3. SOUND RECORDINGS

The sound material of the archive has originally been recorded mainly onto analoguous magnetic tapes. By this method of preservation, however, the recordings cannot be be conserved sufficiently, because the magnetization changes after storing them for some time (the recording will sound damp) or even because the magnetized layer of an old tape will detach from the carrier material. Also, the principal idea of this technique does not allow strictly identical duplicates to be made.
Therefore, the DSAv has been converting sound recordings into a digital formate of archivation since 1994, in order to preserve the valuable materials.
A variation of the Pulse-Code-Modulation-Method has been chosen as a means of convertion: PCM, 44.100 Hz, 16 Bit, mono/stereo with a short RIFF-WAVE- Header. On this method, the saving technique of the worldwide spread audio CDs is based as well. This method conserves all the data produced during the digitization of the analoguous original recording, thus needing a lot of saving capacity. All efficient methods to comprimise audio data, however, are based on deleting information (without without falsifying the physiological listening impression of the original recording "noticably"). This kind of comprimization cannot be completely reversed, so that it is not applicable as a formate of archivation.
The DSAv is, nevertheless, not conceptionally bound to the PCM-formate of archivation. At need, other - especially higher graded - digital recording or saving methods can be integrated into the archive.
In order to physically save the digitized sound recordings from the archive, CD-Rs built up according to the widely spread standard ISO-9660 are used, which can be read by all referant computer systems. The recordings are saved on several CD-Rs for keeping in various places. The DVD-Rs, which originally were dedicated to be used due to their greater saving capacity, are not yet established so far, and thus are not an option for archivation up to now. A "switch" to another carrier for the data with a greater volume doesn't pose a conceptional problem.

2.1.4. ALIGNMENT OF THE SOUND RECORDINGS ANDTRANSCRIPTS

The alignment of sound and transcript, that is, a synchronisation of the sound recordings put into their written form should, following the original concept, entirely be done manually. After wide experiments, however, the alignment in the IDS now succeeded in using a half-automatic method. This method is being optimated for the "less complicated" recordings (about two thirds of the material) at the moment, to be able to use it more efficiently for larger series. The result of the alignment is a list of words including information about the time and length of the words. This time-code is automatically mixed back into the transcript at the corresponding places. They are also ready for archiving without further convertion. They form the base for "phonetic research", where sound extractions around certain words are searched for.

2.2. TOOLS TO BUILD AND USE THE ARCHIVE

For an archive, it is of special importance to keep data and programs conceptionally apart as far as possible, so as not to tie the access to the archived material exclusively to certain programs. Experience shows, that technical possibilities contental necessity concerning the use of data change continuingly. Thus, programs are conceptionally not to be viewed as stable parts. The data material of the archive and their structurs, however, stay unchanged for longer periods of time, so that they have to be prepared to fit changing challenges. That is also why the DV-technical formates have been chosen to allow "switches" to future DV-systems and carrier material without loss.
All the tools for the preparation and use of the material have thus been saved as programming modules. They intercommunicate using only few interfaces and can each be varied or exchanged at need, without majorly influencing other parts of the archive system or even on the data collection. The programs for the preparation and application are mainly based on standard products: The internal SQL-database of the DSAv used for the administration of interactions and material is based on the Interbase-system. It was chosen, because Interbase is satbil, efficient, relatively easy to maintain and accessible for various operating systems. The database tables are developed to be constructed using other SQL-database-systems as well. The accessibility of all database inputs is especially secured by a complete export of all SQL-tables to XML-tables.

2.2.1. RETRIEVAL OF FULL TEXTS

For users, access to the interactions is possible via the documentation as well as via the contents of the transcripts. Thus, a full text retrieval system becomes necessary, since the information in the documentation differ considerably among the corpora and often are not coded strictly. Commonly used Full-Text-Retrieval-Systems could not be used for the purposes of the archive without altering it. The products failed due to their lacking flexibility allowing different types of linguistic questioning to gain acess to the interactions. Thus, the decision concerning the Full-Text-Retrieval tool to applicate finally was made in favour of the system COSMAS II[8], which is being developed for the search within very broad, linguistically annoted corpora of written and

[8] COSMAS stands for "Corpus Storage, Maintenance and Access System".

spoken language.
The DSAv uses COSMAS II neither for "Corpus Storage" nor for "Maintenance", but just as a system for quick access to documentation about interactions or the searched positions within the transcript. The access to COSMAS II-databases has been capsuled overa new client. Thus, COSMAS II-database has for the first time been made accessible by Internet-means out of HTML-masks. Additionally, the complexity of online search language, which makes COSMAS II flexible, could thus be aimfully reduced for the purposes of DSAv transcripts. Like that, the power of the COSMAS II-syntax stays hidden from the user, as long as they don't explicitly want to use it themselves.

2.2.2. ACCESS USING INTRANET AND INTERNET

To make sure, that the access to the material of the archive corresponds to the necessities of Data Security, it has been prepared in four different versions:

- direct, unrestricted access over a fast local net for DSAv employees for the work with the material.
- Staff of the IDS and guest linguists working within the Institute have access to all materials opened for the scientific use. This access use the Intranet of the IDS. After the renovation of the net to state-of-the-art technical standards, the sound material can then be used very efficiently inside the IDS.
- Linguists outside the IDS can gain access to the data opened for scientific purposes as well after applying through the Internet (user name and a password are necessary). Because of general restrictions in the data flow, however, the online access to sound recordings is not totally possible. Here, the quality of sound is reduced in order to quicken the access. People needing the quality of the original recordings have to use the service of the DSAv to have chosen sound material sent to them on CD-R.
- Any interested person can completely anonymously use the documentation about the DSAv-material. Transcripts and sound recordings can, however, only be opened restrictedly.

3. PERSPECTIVES

With the improving technical possibilities of the *world wide web,* the function of internet data bases are becoming more important for data oriented empirical disciplines as linguistics. One important reason for the change from sequential recordings in analogue form is the way of accessing the material. For linguistic purposes an efficient access even to shorter takes out of longer sound recordings are necessary. That is only possible when using digitized recordings saved on carrier materials which are "directly" accessible, without turning reels common in the use of sequential carrier material. Like that, CD-Rs or hard disks are necessary as carrier material and that implies the digitization of the recordings. And finally, only sound recordings in digital formats can be copied for the distant access over data networks.
Up until this point, the lacking clarification and accessibility of this data material has been felt as an essential deficit. But this deficit can be remedied since the digitization of the recordings. The opportunity to edit the sound signal digitally offers a totally new and much easier access to spoken language. Through the integration of the already existing information about the corpora and the transcribed texts in an information- and full text databank, as well as the linking of the data with the acoustic signal (alignment), arises a data-pool with considerably better documentation of the materials and a fast direct grasp of the recorded sounds - internal and external via the Internet. In such a clarified form, the array of the DSAv initiates totally new research questions and prospective for the work at the *Institut für Deutsche Sprache*, as well as for the linguistics altogether.