1. Introduction

1.1 The Collection of German Speech Corpora at the IDS

The "Institut für Deutsche Sprache" (IDS) is hosting a wide range of historical and contemporary German speech corpora. Many of the historical corpora, especially the dialectological corpora like for example the Zwirner Corpus\(^1\) can be accessed online via the "Database for Spoken German" (DGD)\(^2\). Currently we are developing a new, generic speech corpus management system at the IDS where the normalized integration of historical and more recent speech corpora under sustainability aspects and an object-oriented user interface\(^3\) for corpus exploration and querying are monitored as main objectives. The XML schema-based standardization on meta documentation and transcript data level allows the implementation of exact mapping mechanisms for the import and export of existing and future speech corpora. Furthermore, the specific characteristics of individual speech corpora are preserved.

1.2 The Standardization Approach for cross-Corpus Information Management

The theoretical approach of the XML schema based standardization of speech corpus meta data components has been previously described in detail using the example of the corpus instance "German Today"\(^4\).

All speech corpus systems manage meta-information regarding their media source signals. However the information structures of the data components used in different speech corpora may vary considerably regarding the linguistic research questions that are investigated by their creators. Such differences between speech corpora can originate for example from the represented genres, from the degree of content restriction, from the physical data structure or from the research field focussed on (i.e. natural vs. elicited speech)\(^5\).

This article describes our Web-based speech corpus navigation platform that focuses on an abstract standardization concept – matching large speech corpus collections rather than creating particular solutions for data sets of single speech corpus projects. The cross-corpus perspective leads to the definition of a generic, system-wide data model, allowing a smooth corpus data integration without information loss. The components of this data model are hierarchically interlinked with each other as shown in figure 1: the structured XML documentation instances on corpus, event and speaker level, the unstructured, semi-structured or time aligned transcripts, the media sources and perhaps additional unstructured secondary documents are interrelated by system-wide, unique identifiers. The output quality of cross-corpus information processing\(^6\) may vary as it is

---

\(^1\) The Zwirner Corpus was initiated in 1955.  
\(^2\) URL: http://dsav-wiss.ids-mannheim.de/  
\(^6\) like for example the information retrieval in cross-corpus transcript collections
always directly prescribed by the corpus data component with the lowest structure or data quality.

![Components of the Speech Corpus Data Model](image)

Figure 1: Interlinked components of the normalized speech corpus data model

2. The Online Navigation Platform

2.1 The Navigation Interface – Design Principals

Speech corpus navigation interfaces give access to extensive corpus related meta data and transcripts. Nielsen [1994]\(^7\) explains that the structure of graphical interfaces can be guided by either function- or object-oriented design principles. With function-based interfaces, the interface structure is defined by the command set to be implemented which often leads to command-line overloaded interfaces. However the structure of object-oriented interfaces is based on the information units of the documents to be managed by the system. Our speech corpus interface has to provide adaptive views of speech corpus data components. Therefore the document-centric, object-oriented interaction paradigm was chosen to design the application menu and the partitioning of the user’s screen.

The flat structured navigation menu represents the different components of the speech corpus data model (as described in chapter 1.2). It is implemented at the top of the screen with a fixed position in order to permanently provide a system-wide,

homogeneous access to application components. The symbols ▶ and ▼ are indicating whether menu entry points are flat or hierarchically subdivided (cf. figure 2).

Figure 2: The navigation menu of the user interface

For a more intuitive user orientation, the specific types of data components like XML meta-documentations, audio files, transcripts or secondary documents are marked by corresponding classified icons:

Figure 3: The classified icons for specific types of data components

Additionally, "bread crumb" navigation helps the user at any time to identify his current position in the navigation tree and also to navigate back with one mouse click to the previous application level or to the system start page. Figure 4 shows the current location in the navigation tree ("DGD 2.0 intern > Metadaten > Korpusbene > Korpusbeschreibungen") including the two links ("Metadaten" and "DGD 2.0 intern") for back navigation to the second or the first menu level. A large working area is implemented in the centre of the screen to respond to the need to display large meta-documentations and transcripts.

Figure 4: "Bread crumb" navigation and working area
2.2 The Visualization and Presentation of Speech Corpus Content

2.2.1 Generic Visualization of the XML Meta-Information of Speech Corpora

All speech corpus related XML meta-documentations are stored natively and valid against the corresponding XML catalogue schema in the system’s XML database. To avoid corpus specific instance visualizations, a generic XML document rendering module has been implemented. It provides XML tree visualization, expandable / collapsible document nodes and a node level selection functionality. Hyperlinks to external resources are directly accessible from the XML visualization (cf. figure 5).

The cross-corpus display method of corpus-, event- and speaker documentation offers an ergonomic navigation experience especially for large data-centric XML documentation instances.

Speech corpus projects also may document the geographic coordinates of their event locations like for example in the case of the project "German Today". A geographic map can be displayed on demand for each event. Figure 6 shows the geographic map for the event DH--_E_00167 (with geographic latitude 47.423336 and longitude 9.377225) which took place in St. Gallen (Switzerland).

---

8Cf.: http://www.ids-mannheim.de/prag/AusVar/Deutsch_heute/
2.2.2 Transcript Visualization and Presentation

The more speech corpora (and especially historical corpora) belong to the corpus collection, the more a common concept of "transcript" becomes fuzzy – in the sense of the annotated phenomena and first of all in the sense of the data formats used. Recent speech corpus projects mostly use one of the annotation tools available nowadays\(^9\). These tools clearly separate annotation content from its graphical representation and most of them also include time alignment. However, the transcripts of historical corpora consist of project-specific data formats, often exclusively oriented on the graphical layout of annotation content without a time alignment of the annotated information. Therefore our speech corpus management system has to operate with unstructured, purely layout-oriented data formats\(^10\) and with transcript editor specific document structures at the same time. At this point, no generic visualization of transcript information is possible. Concepts here mainly depend on the structuring degree of the transcripts.

If we look at the transcript-related part of the meta-documentation, a generic visualization of information is possible. Corpus-specific transcript lists are accessible via the main menu entry point "Transkripte". Figure 7 shows the corpus-specific transcript access list for the speech corpus DS\(^11\).

---

\(^9\) A list (in german language) of currently available annotation tools can be found at the following URL: http://prowiki.ids-mannheim.de/bin/view/GAIS/TranskriptionEditoren

\(^10\) like for example text, html or proprietary word processor formats

\(^11\) DS: Dialogstrukturen
2.2.3 Media Presentation

Speech corpora may include different types of interdependent media files. First, each corpus event is related to one or more source files. This is the raw material of the recorded event, originating for example directly from an audio recording device. As an event may be composed of several speech events, the original raw material can be further segmented into speech event specific recordings. All relevant information regarding the different media file types is maintained in the meta-documentation of the corresponding event and can be accessed via the list of the main menu entry point "Aufnahmen" as shown in figure 8.
3. Retrieval Strategies for unstructured and structured Speech Corpus Data Components

Media file content cannot be located without descriptive meta-information\(^\text{12}\). Current speech corpus retrieval systems gather information about media source files by exploring the corresponding meta data and possibly related transcript data. As described in the previous chapter, the transcript data components of speech corpus collections may spread regarding their structuring degree. Appropriate retrieval strategies mainly depend on this degree. For structured transcripts the distinction between a document-centric and a data-centric organization of the annotation layers is also of interest\(^\text{13}\). Document-centric (single-layer) transcripts can be seen as semi-structured transcripts, enabling only full-text search results that are enhanced by additional time alignment information. This also means that a complex layer-aware query processing against transcripts with one specific multi-layer structure will have to be downgraded to a simple full-text search as soon as different (document-centric) structures from other corpora join the same query. In a worst case scenario, the same partial incompatibility effect can even occur with transcripts of one single corpus that were created with a current multi-layer transcript editor. Different reasons may have contributed to inhomogeneous transcript data:

- Signal segmentation without precise segmentation guidelines (i.e. phones, words, phrases or turns)
- No or not sufficient naming conventions applied for the different transcript layer descriptors (i.e. no unique descriptor used for orthographic transcription layer)
- No exact semantic layer definition available or semantic mix-up of layer content (i.e. mix-up of orthographic and phonetic markup in one single layer)
- No exact syntactic definition of layer content available or syntactic mix-up of layer content (i.e. mix-up of punctuation- or capitalization conventions in the orthographic layer)
- Violation of cross-layer time relations (i.e. caused by interval changes that were made with multi-layer transcript editors without layer inheritance control)

All these reasons may cause structural incompatibilities with the result that full-text search is the lowest common denominator regarding feasible retrieval strategies. In an online environment, the determination of the appropriate retrieval strategy has to be an automated process depending on the user's current corpus selection.

3.1 The Full-Text Search Module

The implementation of full-text search functionality does not require structured data. But at the same time structured data can also be included during the search. For semi- or unstructured transcript data only a full-text search interface can be provided. As we already use the Oracle XML Database for our XML meta-documentation storage needs, we can also benefit from the rich full-text search features provided by Oracle Text\(^\text{14}\) for

\(^{12}\) Cf. U.S. General Services Administration (2005), p. 9
\(^{13}\) Cf. Trippel (2004), p. 2
\(^{14}\) Cf. Schneider (2009), p. 144-147, (in german language)
the full-text search in transcripts. Here some examples of the full-text query features provided:

- The simple and multiple wildcard characters "_" and "%":
  
  - _ind matches i.e. "Kind" and "Wind"
  
  - %wind matches i.e. "Nordwind" or "Südwind"

- The operators AND and OR build logical relations between search terms:
  
  - Nordwind AND Südwind matches only documents with occurrences of both terms
  
  - the NOT operator excludes a specific search term:
    Nordwind NOT Südwind matches only documents containing "Nordwind" but not containing "Südwind"

- The NEAR operator finds documents depending on the word distance of search terms:
  
  - NEAR(Schule, Kirche, 4, true) matches documents where both search terms occur with a (maximum) word distance of 4 words.

A working application example is shown in figure 9 (the audio segments corresponding to the query hits are directly accessible by pressing the corresponding button).

---

Figure 9: Full-text search in semi-structured transcript data with search results (KWIC-list)

Context sensitive straight to the point queries as in the case of XQuery interfaces are not possible with full-text search against unstructured data. Short query response times and user interfaces that are simple to handle are the main advantages of full-text retrieval strategies.

However, an additional full-text option for structured documents (like speech corpus XML meta-documentation) can also be helpful for users that are not familiar with the specific document structure. The disadvantages are obvious: the hits provided by full-text search in structured documents are completely context-free because an exact specification of document sections (like single or complex XML elements) as with the use of XQuery is not possible.
3.2 XQuery Information Retrieval in structured XML Documents

For the information retrieval in fine-grained XML instances like meta data or time aligned multi-dimensional transcripts, the full-text search option without context consideration may not be sufficient. For the hierarchically interdependent informational units of XML structured data a context sensitive retrieval approach is implemented based on the XML query language (XQuery). XQuery allows the definition of complex queries, providing a wide set of XML based retrieval capabilities like for example criteria specific information selection and filtering, joining of data from document selections, sorting, grouping, aggregating, transforming and restructuring of data and arithmetic calculations on numbers and dates\textsuperscript{15}.

XQuery is a powerful query language operating on XML instances stored in an object-relational XML database, similar as SQL\textsuperscript{16} queries are working against relational database tables. In both query cases, a deeper knowledge about the underlying information structure is required for the person who is defining and executing the query. This means for Web-based XQuery retrieval interfaces: two different approaches can be implemented for online XQuery execution.

The first is a HTML form providing a graphical representation of the XML information tree including input fields or listings with possible field contents. After submitting the form, the XQuery syntax code is generated from the HTML form inputs and executed on server side. Figure 10 shows a draft of a graphical XQuery composition interface:

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{xquery_interface.png}
\caption{HTML form providing a graphical XQuery composition interface}
\end{figure}

\textsuperscript{16} SQL: the Standard Query Language (relational databases)
XQuery processing via HTML form interfaces may be limited regarding query definition flexibility as it quickly becomes costly in terms of programming effort, especially with queries operating on complex data-centric XML instances or with queries joining distinct XML structures.

A less graphical approach where the XQuery can be edited manually (cf. figure 11) is more suitable to query this type of XML instances. The HTML form provides a text area field where the XQuery syntax can be freely entered as plain text. When the form is submitted, the XQuery input string is directly executed on the database server. This approach is primarily intended for system experts.

Figure 11: HTML form for XQuery plain text submission

In the above expert XQuery example, we look for all speakers of the speech corpus "German Today"\textsuperscript{17}, currently living in Germany with at least one parent coming from Austria. The query matches one single speaker (id DH--S\_00155) whose father was born in Austria.

4. Summary and Outlook

Media source files become analyzable via their appropriate meta-information. Contemporary speech corpus systems have to close the gap between the processing of binary media data and related meta-information. The need for standardization of speech corpus components is commonly accepted. Nevertheless, the identification of all parameters necessary for a cross-corpus standardization still remains an outstanding goal. Future evolving technologies like the MPEG-7 standard\textsuperscript{18} might provide appropriate logic to achieve the standardized integration of the different audiovisual

\textsuperscript{17} The speech corpus project "German Today" (corpus id DH) is ongoing. Currently about 600 speaker documentation instances are stored in the XML database.

\textsuperscript{18} MPEG-7 ISO/IEC Standard
information types like audio, voice, video, images, graphs and 3D models, potentially involved in media corpora.

References


Table of Contents

1. Introduction ........................................................................................................................................... 1
   1.1 The Collection of German Speech Corpora at the IDS ................................................................. 1
   1.2 The Standardization Approach for cross-Corpus Information Management ............................ 1
2. The Online Navigation Platform ............................................................................................................ 2
   2.1 The Navigation Interface – Design Principals ............................................................................. 2
   2.2 The Visualization and Presentation of Speech Corpus Content ............................................... 4
       2.2.1 Generic Visualization of the XML Meta-Information of Speech Corpora .................... 4
       2.2.2 Transcript Visualization and Presentation ........................................................................... 5
       2.2.3 Media Presentation ................................................................................................................. 6
3. Retrieval Strategies for unstructured and structured Speech Corpus Data Components .................. 7
   3.1 The Full-Text Search Module ........................................................................................................... 7
   3.2 XQuery Information Retrieval in structured XML Documents .................................................. 9
4. Summary and Outlook ............................................................................................................................. 10
References .................................................................................................................................................. 11

Table of Figures

Figure 1: Interlinked components of the normalized speech corpus data model ............................. 2
Figure 2: The navigation menu of the user interface ............................................................................. 3
Figure 3: The classified icons for specific types of data components ............................................ 3
Figure 4: "Bread crumb" navigation and working area ........................................................................... 3
Figure 5: Generic XML document rendering ......................................................................................... 4
Figure 6: Geographic map (based on documented geocodes showing the location of an event) ......... 5
Figure 7: Corpus-specific transcript list for the speech corpus DS .................................................... 6
Figure 8: Corpus-specific list of source recordings for the speech corpus DH .................................. 6
Figure 9: Full-text search in semi-structured transcript data with search results (KWIC-list) ........... 8
Figure 10: HTML form providing a graphical XQuery composition interface ............................... 9
Figure 11: HTML form for XQuery plain text submission ................................................................. 10