

## Querying Spoken Language Corpora

Thomas Schmidt, IDS Mannheim ([thomas.schmidt@ids-mannheim.de](mailto:thomas.schmidt@ids-mannheim.de))

Spoken language, as recorded in an audio or video file, is made accessible for systematic corpus linguistic queries through an appropriate transcription. The TEI Guidelines provide means of representing such transcriptions in XML documents, parts of which – for instance: words and sequences of words – can be treated in a largely analogous manner to TEI encoded written language.

In my contribution to the workshop, I will focus on challenges which arise where that analogy ceases to hold. Two aspects are central in this respect:

First, working with spoken language requires an extended notion of context. In corpora of written text, the context of an item usually means the immediately preceding and following words and some global metadata (author, time, etc.) pertaining to a text as a whole. By contrast, a query on spoken language corpora must take into account additional types of context, such as actions and words which are parallel to the item in question and metadata which pertains to a part of the “text” only (such as properties of the speaker of an individual utterance). Moreover, since transcriptions are reduced representations of the actual linguistic object, it is often necessary to identify the context of an item also in the original recording. Query mechanisms for spoken language have to make sure that these additional types of context are available to the user.

Second, spoken language transcriptions have a more complex internal structure than their written text counterparts. Not only do they contain tokens which are neither words nor punctuation (for instance: pauses, descriptions of non-verbal activity, incomprehensible material), but they also have to represent the interactional structure of a speech event with speaker changes and simultaneity and overlap of utterances. This additional structural complexity has to be taken into account when query strategies for written text are transferred to spoken language corpora. Furthermore, the structural properties can also be the object of a query in itself, for instance when we want to find instances of a given item in the vicinity of a pause, near a speaker change or inside an overlap.

I will start my contribution by giving an overview of these two types of challenges and explaining what data models and data formats can be used to address them and how they relate to the TEI guidelines (Schmidt 2011). I will then present two query environments – the EXAKT tool of the EXMARaLDA system (Schmidt/Wörner 2013, <http://www.exmaralda.org>) and the web interface of the Database of Spoken German (DGD2, Schmidt, Dickgießer & Gasch 2013, <http://dgd.ids-mannheim.de>) – in which some (but by no means all) of the requirements for querying spoken language corpora are implemented. Most importantly, both systems have query interfaces which

- allow the user to access all the different types of context that can be relevant for a query,
- give the user various possibilities to build a complex query out of smaller and simpler components, and
- enable the user to manually interact with query results, thus making it possible to discard false positives which cannot be identified via a formal query

Experience so far shows that users are indeed able to work productively with spoken language corpora in these environments. It remains an open question, however, how well the environments

adapt to additional query requirements, to larger and possibly more heterogeneous bodies of data, and – maybe most importantly – what possibilities there are to integrate them into more generic query environments, for instance via a standardized query language in a federated search environment. I will conclude my contribution with a discussion of these open questions.

### References

**Schmidt, T. (2011)** A TEI-based Approach to Standardising Spoken Language Transcription. In: Journal of the Text Encoding Initiative (1). [<http://jtei.revues.org/142>]

**Schmidt, T. & Wörner, K. (2013)** EXMARaLDA. To appear in: Jacques Durand, Gut, Ulrike & Kristoffersen, G. (ed.): Handbook on Corpus Phonology, Oxford University Press.

**Schmidt, T. & Dickgießer, S. & Gasch, J. (2013)** Die Datenbank für Gesprochenes Deutsch - DGD2. Submitted to: Gesprächsforschung Online.