Accepted | May 2015 | Comments Welcome | sauersim@hu-berlin.de | anke.luedeling@hu-berlin.de

# Flexible Multi-Layer Spoken Dialogue Corpora

Simon Sauer and Anke Lüdeling (Humboldt-Universität zu Berlin)

***Abstract***

This paper describes the construction of deeply annotated spoken dialogue corpora. To ensure a maximum of flexibility – in the degree of normalization, the types and formats of annotations, the possibilities for modifying and extending the corpus, or the use for research questions not originally anticipated – we propose a flexible multi-layer standoff architecture. We also take a closer look at the interoperability of tools and formats compatible with such an architecture. Free access to the corpus data through corpus queries, visualizations, and downloads – including documentation, metadata, and the original recordings – enables transparency, verifiability, and reproducibility of every step of interpretation throughout corpus construction and of any research findings obtained from this data.

***Key words***

Spoken corpora, multi-layer architecture, standoff, annotation, annotation tools, search, visualization, BeMaTaC

## 1. Introduction

How do learners use prosody in connection with information structure? How does gender influence the use of filled pauses? How many utterances in spontaneous dialogue occur without a finite verb? How do people conceptualize space? How does backchanneling work in a task-oriented setting? How do speakers converge or diverge over the course of a conversation?

This random selection of questions shows that a spoken corpus can be used for many different purposes and that not all research questions are known at the time of

corpus construction. Each research question demands a different categorization and interpretation of the same primary data (for a discussion of what counts as primary data, see Section 3). Ideally, the interpretation is coded within the corpus as annotation. While this may sound trivial, many corpora are constructed for a very specific research question and released only with annotations relevant to that question – and while we focus on spoken corpora here this is true for most written corpora as well. Even for reference corpora that contain different spoken genres and are specifically designed to be used for different research questions such as the BNC (Burnard 2007), the user cannot easily add annotation layers within the corpus, while keeping all existing segmentations and annotations. Spoken corpora are expensive to construct and a well-designed corpus can typically be used by researchers other than the original constructors for purposes and questions that extend the original project.

This paper describes the construction of flexible multi-layer spoken dialogue corpora that can be modified and extended at any stage for applications that were not envisaged at the time of corpus construction. The focus of this article lies on architecture and annotation rather than on the sampling or recording of primary data or on a specific corpus. We will occasionally illustrate our approach using BeMaTaC, a German map-task corpus (briefly introduced in Section 2), but everything we say about best practice is independent of this specific corpus. In Section 3.1, the paper will then discuss the architectural requirements essential for a linguistically annotated spoken corpus before going into the details of multi-layer architectures in general and multiple tokenizations in particular in Section 3.2. Section 4.1 will deal with tools for automatic processing and manual annotation, while Section 4.2 is concerned with format incompatibilities and feasible solutions for such problems. Finally, we will discuss

different aspects of corpus access, namely search (Section 5.1), visualization (5.2) and the data available for download (5.3).

## 2. BeMaTaC

BeMaTaC, the Berlin Map Task Corpus, is a freely available and deeply annotated multimodal map-task corpus of spoken learner and native German. It uses a map-task design, where one speaker (the instructor) instructs another speaker (the instructee) to reproduce a route on a map with landmarks (Anderson et al. 1991).[1]

The dialogues are recorded with two separately placed microphones (only one of the recordings will be used for further processing) and a video showing the drawing hand of the instructee. Transcriptions are consistently tokenized, time-aligned, separately normalized and annotated on a wide range of different layers. Annotations include e.g. part-of-speech tags, utterance spans, backchanneling, disfluencies and repairs, as well as syntactic dependencies (see Section 4.1 for details). Extensive and anonymized metadata are provided with every dialogue. For more information about the corpus, see BeMaTaC (2014), Giesel et al. (2013) and Sauer & Rasskazova (2014).

## 3. Architecture

### 3.1 Primary data and annotations

What constitutes the primary data in a linguistic investigation is not always obvious and has been discussed extensively (see e.g. Himmelmann 2012 for a general overview, Wichmann 2008 for a discussion of primary data in spoken corpora and Kirk, this volume). In corpus linguistics, this is especially difficult for those types of data that have to be rendered into some kind of symbolic sequence in order to be used in research, such as auditory or visual data. Here, primary data is often understood to refer to the original sound or video recordings. The majority of linguistic analysis and

annotation, however, is based solely on a transcription of the audio signal (or on a transliteration of the video signal, as in sign language or gestural research, see e.g. Hanke & Storz 2008). Any transcription, of course, is an interpretation of the data and results in a loss of information, as Thompson (2005) and many others have pointed out. This is unavoidable and true for any mode of transcription (narrow, broad, orthographic, etc.). In our corpus model, we call the transcription primary, as all layers of annotation and analysis reference this layer. The recordings of spoken data therefore have to be modelled as annotation (this is in contrast to the actual technical implementation as detailed in Section 3.2).

The primary data needs to be split up into basic units (tokens). Tokens are the smallest unit that can be annotated. Tokens are often used to indicate corpus size and as a normalizing base for corpus counts. For written corpora of most European languages, the tokens are graphemic words (typically sequences of characters between whitespaces)[2]. For spoken corpora, however, there is no agreement on what constitutes a token. We find spoken corpora with utterances, turns, elementary discourse units, or sometimes ‘sentences’ (we will not enter into the discussion of what constitutes a sentence in spoken language) as smallest units. A larger basic unit is not useful for research questions pertaining to smaller units. For example, the distribution of part-of-speech tags can be used to compare spoken and written registers. However, part-of-speech tags cannot be assigned if the smallest unit that can be annotated is an utterance. The opposite – namely annotating a larger unit – is always possible by annotating a sequence of tokens, in a so-called span annotation.

We therefore argue that spoken corpora should be tokenized consistently and similarly to written corpora. This way, we can annotate any unit we want and compare

corpus counts statistically. Just like the transcription itself, the tokenization is an interpretation of the data.

As transcription and tokenization are interpretations, we will show in Section 3.2 that a flexible corpus architecture should allow for different transcriptions and different tokenizations of the same sound file (in essence, this means that it must be possible to have different primary layers).

After tokenization, a corpus needs to be annotated. The following sections deal with different aspects of annotation; here we want to briefly introduce the basic formats so that we can refer to them later (see e.g. Carletta et al. 2003 or Chiarcos et al. 2009):

(a) Token-based annotation, where a category (tag) refers to a given token. Some token-based annotation layers assign a tag to each token (such as part-of-speech layers), some assign a tag only to certain tokens (such as disfluency layers, see below).

(b) Span-based annotation, where a tag is assigned to a range of tokens (such as multi-word units or elementary discourse units).

Token and span-based annotations are unstructured as such. In addition to such 'flat' annotation layers, it is possible to have more complex layers:

(c) Hierarchical annotation, where a tree or graph is assigned to a span of tokens. This could be a syntactic tree or graph (Nivre 2008) or a tree pertaining to discourse features such as rhetorical structure or argument structure (see e.g. Stede 2011 for an overview of text and discourse annotations in corpora).

(d) Pointing relations, where a token points to one or more tokens (such as in anaphoric relations or co-reference annotations).

With these basic formats, it should be possible to add information for all relevant research questions. It is obvious that a corpus architecture should be flexible enough to permit all those different abstract annotation formats independent of their specific content. Multi-layer models allow exactly that, which we will illustrate in detail in the following section.

### 3.2 Multi-layer architectures and multiple tokenizations

Roughly speaking, multi-layer (standoff) models store all annotation layers separately from each other and from the primary data (Carletta et al. 2003, 2005),[3] which means that it is possible to add as many annotation layers as needed and that new annotation layers do not interfere with already existing annotation layers. We use a very flexible and powerful multi-layer model (Chiarcos et al. 2009, Zeldes et al. 2009) which puts (almost) no restrictions on the formats of the primary data and the formats of the annotations (see Krause et al. 2012, Krause & Zeldes, 2014). Annotation layers can be visualized, edited, added, removed, and searched independently, thus also allowing overlapping annotations.

We want to illustrate the need for flexible multi-layer models in dealing with spoken corpora by addressing two common problems: overlapping segments and different transcriptions.

As stated above, many corpus models break up the primary data into tokens. Tokens are used in search and in visualization: Many search tools allow the user to specify conditions over a token (e.g.: it must be a verb and begin with *be*-). In the visualization of the search results the user can often specify a left or right context of *n* tokens. Token-based models, however, often have problems in dealing with overlapping segments in dialogue because there can only be one layer of tokenization. Figure 1

illustrates how the instructor (upper transcription) is still uttering a filler, while the instructee takes advantage of the hesitation and already starts with his utterance.

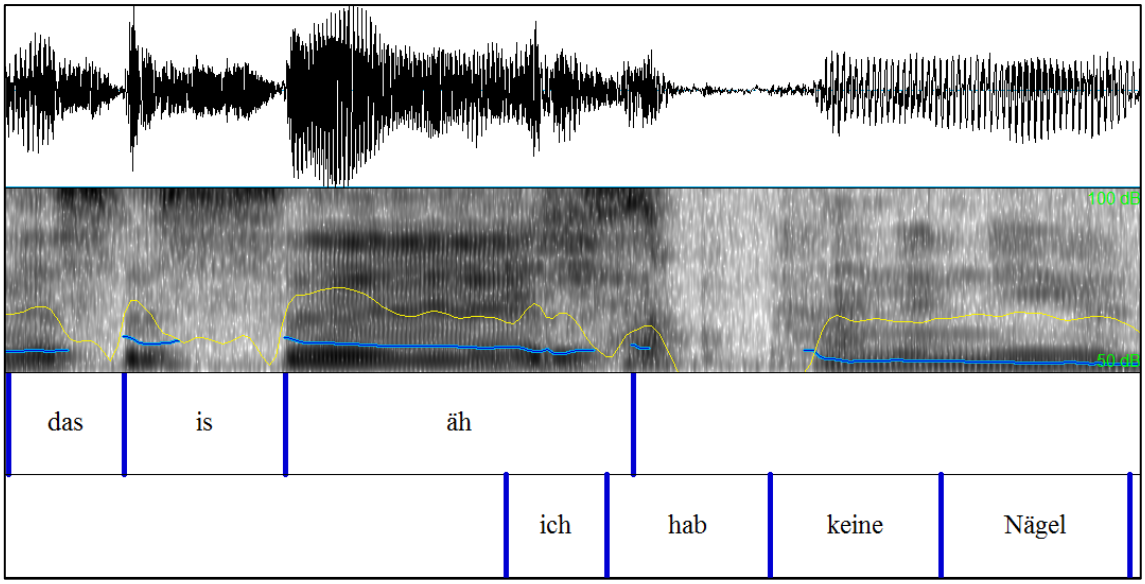


*Figure 1: Overlapping speech in BeMaTaC (subcorpus L1, version 2013-02, document 2011-12-14-A, token range 270 to 277): 'this is er' – 'I have no nails'*

Moreover, in order to allow speaker-specific analyses and visualizations, every single utterance would have to be annotated with its corresponding speaker, resulting in a massive overhead. Annotating segments made up of multiple tokens, such as sentence spans containing word-level tokens, becomes impossible when a sentence exceeds one uninterrupted utterance, as can be seen in Figure 2. A separate transcription layer for each speaker is necessary and these layers must be completely independent from each other, as they will have conflicting tokenizations. As some tokens will only exist in one speaker's transcription, others, however, only in another speaker's transcription, a minimal tokenization does no longer exist.

| **instructor** | dann | legen | wir | mal | los | mit | der | | Beschreibung |
|---|---|---|---|---|---|---|---|---|---|
| **utterances** | utt | | | | | | | | |
| **instructee** | | | | | | | | jo | |
| **utterances** | | | | | | | | utt | |
| **pauses** | | | | | | | | | 0.2 |

*Figure 2: Interrupted utterance span and independent pauses layer in BeMaTaC (subcorpus L1, version 2013-02, document 2011-12-14-A, token range 1 to 10): 'then let us get started with the' – 'yeah' – 'description'*

Dialogues are therefore often stored in timeline models. This means that there is a real or abstract 'timeline' to which each signal refers. In a sense, the timeline is the primary data and the transcriptions are annotations (Schmidt 2004, Wörner 2009; note that this is a technical operationalization only and therefore in contrast to the theoretical model as described in Section 3.1). The timeline can also be used to link to a position in an audio or video file, if the data is fully aligned with the original recordings. Whereas timeline-based models seem very 'natural' for spoken data and are very good at handling overlapping segments, they sometimes have problems in defining tokens for the purpose of search and visualization. Krause & Zeldes (2014) describe the model we are proposing. It combines the advantages of both token-based and timeline-based models: The primary data is segmented into tokens. To these, the model adds a layer with information about the precedence of the tokens, which can be understood as an artificial timeline. This layer can be used in search and visualization (for the technical details, see Krause & Zeldes, 2014).

Many papers about spoken corpora debate the degree of normalization in the chosen transcription (anything between a narrow phonetic transcription to a fully normalized orthographic transcription, see e.g. Thompson 2005). Choosing a narrower transcription will inevitably result in a higher degree of variation and less consistency, complicating corpus queries and automated processing such as part-of-speech tagging. A higher degree of normalization will again lead to a loss of information. In 'traditional' corpus architecture, this is indeed a problem because the transcription is the primary data for all annotations. In our architecture it is possible to have different 'primary'

layers. Moreover, a colloquial contraction like *gehste* ‘go-you’ does not have its own part-of-speech tag, but by means of an additional ‘primary’ layer with a normalized two-token *gehst du*, the two separate part-of-speech tags can simply be mapped to this layer.

Another matter of debate is the treatment of pauses: In a traditional token-based model, pauses must be part of the preceding or of the following utterance and, consequently, one speaker is represented as being responsible for the pause. This is unsatisfactory, as a pause is simply a period of silence, for which all speakers are equally responsible. This can be modeled in a speaker-independent layer, independent of the transcription proper. An example of this can be seen in Figure 2. Similar issues arise with non-verbal sounds such as laughter or coughing. While they can be attributed to a speaker, they are independent of the primary (verbal) transcription.

## 4. Tools and formats

### 4.1 Tools

In this section we want to give another reason for using a general multi-layer standoff model for spoken corpora: the necessity for using different annotation tools on the same primary data.

Some annotation layers must be added manually, some can be constructed semi-automatically, and others can be produced fully automatically. There are different tools for different types of annotations. For most purposes or types of annotation one can find dedicated tools, such as Praat (Boersma 2010), EXMARaLDA (Schmidt & Wörner 2009), or ELAN (Sloetjes & Wittenburg 2008) for the annotation of spoken data, TrED (Pajas & Stepanek 2008) or Arborator (Gerdes 2014) for dependency annotation, and MMAX (Müller & Strube 2006) for coreference annotation. A number of multi-purpose

annotation tools such as brat (Stenetorp et al. 2012), Atomic (Druskat et al. 2014), and WebAnno (Yimam et al. 2013) are under development but so far have severe limitations when it comes to spoken data. In addition to the different functionalities it is sometimes just a matter of personal or community-wide preference that leads to the choice of one tool over another.

We want to illustrate the need for different tools and procedures with a concrete example taken from BeMaTaC. Our initial transcription is done word by word – ignoring speaker overlaps and without any form of audio alignment. This can therefore be accomplished with a simple text editor and an audio player software supporting global keyboard shortcuts – or with any setup that provides plain text. Our excerpt will then look like in Figure 3.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| an | seiner | rechten | Seite | ja | und | dann | schlägste | ähm | schä/ | schlägst | du | Sozusagen | | |
| on | his | right | side | yes | and | then | turn-you | erm | tr/ | turn | you | so-to-speak | | |

*Figure 3: Initial basic transcription in BeMaTaC (subcorpus L1, version 2013-02, document 2011-12-14-A, token range 640 to 656): The slash marks an intra-word truncation.*

Next, we automatically process these transcripts with MAUS (Schiel et al. 2011), a tool that automatically segments and aligns the transcript with the audio recordings, both on a sound and word level, as can be seen in Figure 4. While this process may not always be perfectly accurate, it provides an alignment that is both systematic and predictable.

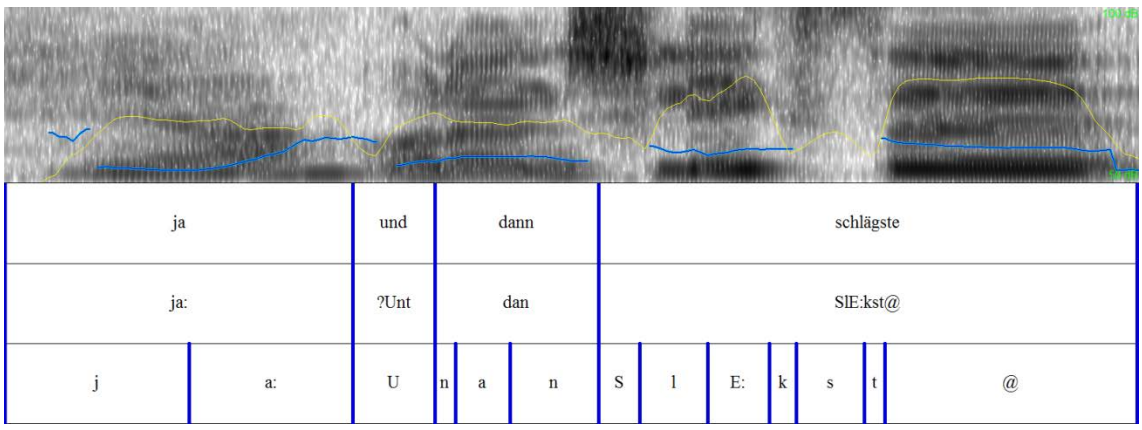


*Figure 4: MAUS segmentation output in Praat.*

We use Praat, which is primarily designed for phoneticians and thus allows a very fine-tuned alignment, to then correct any misplaced token boundaries and to divide the transcript into two separate layers, one for each speaker (compare the location of *ja* 'yes' in Figure 4 with its proper speaker assignment in Figure 5, which shows the final result in EXMARaLDA). In order to have separate layers for the normalization, the transcription layers are then duplicated and orthographically normalized if tokens need to be split up, in our example when the contraction *schlägste* becomes two separate tokens *schlägst* and *du* (see lines 2 and 3 in Figure 5). Moreover, audible extra-lingual events such as laughter are marked on separate layers (marked as *instructor [extra]* and *instructee [extra]* in Figure 5).

As soon as the alignment is fixed, any further normalization is continued in the EXMARaLDA Partitur Editor, which focuses less on audio recordings and more on transcription and annotation. It enables easier handling of large numbers of different layers and provides simple mechanisms to merge tokens for span annotations and to split them again. Data subsequently added in EXMARaLDA includes dialogue and speaker metadata as well as a variety of annotations such as syntactically motivated utterance spans (in the layers marked *instructor [utt]* and *instructee [utt]* in Figure 5), backchanneling (*bc*), disfluencies (*df*) and repairs (*repair* and *subrep*; see Belz 2013 for

| [tok] | 640 | 641 | 642 | 643 | 644 | 645 | 646 | 647 | 648 | 649 | 650 | 651 | 652 | 653 | 654 | 655 | 656 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| instructor [dipl] | an | seiner | rechten | | Seite | | | und | dann | schlägste | | ähm | | schä/ | schlägst | du | sozusagen |
| instructor [norm] | an | seiner | rechten | | Seite | | | und | dann | schlägst | du | | | | schlägst | du | sozusagen |
| instructor [lemma] | an | sein | recht\|rechts | | Seite | | | und | dann | schlagen | du | | | | schlagen | du | sozusagen |
| instructor [pos] | APPR | PPOSAT | ADJA | | NN | | | KON | ADV | VVFIN | PPER | | | | VVFIN | PPER | ADV |
| instructor [utt] | | | | | | | | utt | | | | | | | | | |
| instructor [df] | | | | | | | | | | pr | | fl | | | | | |
| instructor [repair] | rs | | | | | | | | | rd | | ir | | rs | | | |
| instructor [subrep] | rl | sl | il | | i2 | | | | | | | | | sl | s2 | il | |
| instructor [repair2] | | | | | | | | | | | | | | rd | rs | | |
| instructor [subrep2] | | | | | | | | | | | | | | | sl | | |
| instructor [extra] | | | | lacht | | | | | | | | | | | | | |
| instructee [dipl] | | | | | | | ja | | | | | | | | | | |
| instructee [norm] | | | | | | | ja | | | | | | | | | | |
| instructee [lemma] | | | | | | | ja | | | | | | | | | | |
| instructee [pos] | | | | | | | ADV | | | | | | | | | | |
| instructee [utt] | | | | | | | utt | | | | | | | | | | |
| instructee [bc] | | | | | | | bc | | | | | | | | | | |
| instructee [df] | | | | | | | | | | | | | | | | | |
| instructee [repair] | | | | | | | | | | | | | | | | | |
| instructee [subrep] | | | | | | | | | | | | | | | | | |
| instructee [repair2] | | | | | | | | | | | | | | | | | |
| instructee [subrep2] | | | | | | | | | | | | | | | | | |
| instructee [extra] | | | | | | | | | | | | | | | | | |

*Figure 5: EXMARaLDA Partitur Editor with various token and span annotations.*

an overview of repair categories). Some annotations rely on information only contained in the audio. In this example, the *pr* tag in the instructor's disfluencies layer (*df*) represents a prolongation. In cases like these, the tool of choice should support immediate playback of the corresponding audio segments for the token currently being annotated.

Part-of-speech tags (*pos*) and lemmatization (*lemma*) are automatically generated with TreeTagger (Schmid 1994), using the Stuttgart-Tübingen-TagSet (Schiller et al. 1999).[4] Scripts are used to further add simple information such as a token numbering (*tok*), token length (*len*), or periods of silence (*break*). A subset of BeMaTaC is part of the NoSta-D corpus (Dipper et al. 2013) and features named entities as well as

dependency relations and co-reference chains, which have been annotated with WebAnno, as shown in Figure 6.

### 4.2 Formats and standards

The examples in Section 4.1 illustrate that it is sometimes useful to use several tools to annotate different properties of the corpus. However, as soon as more than one tool is

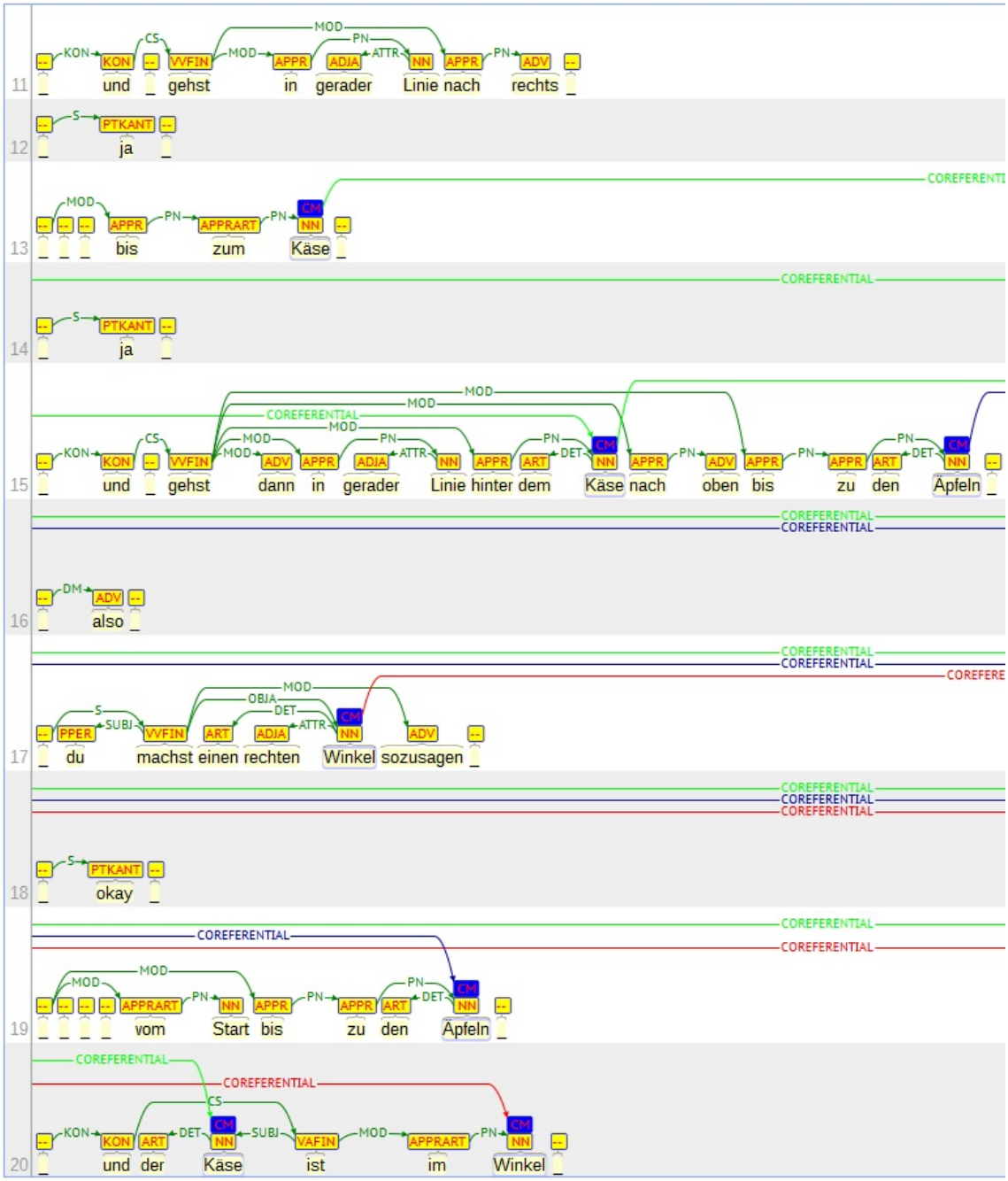


*Figure 6: Coreferences and dependencies in WebAnno (NoSta-D, subcorpus BeMaTaC, document 2011-12-14-B).*

employed, interoperability issues will arise: Each tool requires a given input format and produces a specific output format. The task then is to 'chain' or convert these formats in a way that makes it possible to combine the different annotation layers. This has led to the construction of toolchains such as WebLicht (Hinrichs et al. 2010). These offer a range of flexibly combinable processing tools, each converting the previous tool's output to a common format (TCF in WebLicht) that can be converted to an input for the next tool. To date, however, these environments do not support spoken data.

Some tools support import and export in various formats but there obviously is a practical limit. In the BeMaTaC pipeline illustrated above, MAUS offers a built-in function to export into Praat's proprietary TextGrid format, which is not based on XML but nonetheless text-based and self-descriptive. EXMARaLDA in turn supports the import of Praat TextGrids. TreeTagger output, which uses a plain-text tab-separated values format, can be imported into EXMARaLDA – but not into existing time-aligned data. The accessibility of both formats, however, allows this gap to be filled with automatic conversion scripts.

There is no ubiquitous and universally accepted standard – and there probably never will be. Even widely-used formats such as those proposed by the TEI (TEI Consortium 2014) do not support every type of annotation and architecture.[5] Moreover, they evolve and different researchers have very different needs and expectations. As tools do not and cannot support every format and standard, there is a need for format diversity – at least to a certain degree. However, $n$ different formats will result in a potential need of $n^2 - n$ different mappings.

This is where converter frameworks such as SaltNPepper (Zipser & Romary 2010) come in. Salt is an abstract graph-based linguistic meta-model which works as an

intermediary between different formats[6]. The conversion framework Pepper is based on this meta-model and provides a plug-in framework for an unlimited number of modules, importing or exporting data to or from Salt. The modularity enables the combination of existing modules, handling already supported formats, and new formats. This reduces the total number of necessary mappings to 2*n*.

SaltNPepper already provides support for numerous formats such as CoNLL (Buchholz & Marsi 2006), the ISO-standardized GrAF (Ide & Suderman 2007), PAULA (Dipper 2005) or generic XML. In BeMaTaC, this allows us to import all our EXMARaLDA data, including metadata and aligned audio/video files, use the built-in timeline to automatically create a minimal token layer and export everything into relANNIS, the database format of our search and visualization architecture ANNIS (Zeldes et al. 2009).

## 5. Access

Spoken corpora are particularly time-intensive and thus expensive to build, for this reason it is our firm conviction that they should be made as easily accessible as possible, thus alleviating the scarcity of spoken corpora openly available for research.[7] Limited access to a corpus means limited reproducibility, verifiability and improvement of any findings thus obtained. We propose licensing under a Creative Commons license (Creative Commons 2014), which allows a variety of different configurations. BeMaTaC is published under a Creative Commons Attribution license, or CC-BY in short, thereby not restricting usage of the data e.g. in commercial journals – which may be impeded by using non-commercial or share-alike restrictions – while at the same time ensuring proper attribution and preventing abuse.

We propose four main functionalities that should cover both one's own research but also access for other researchers with fundamentally different research questions: It should be possible to (a) search the corpus and interpret the results in a qualitative manner, (b) search the corpus and interpret the results quantitatively and statistically, (c) download all or parts of the corpus, and (d) modify or extend the corpus.

The two primary means of accessing data are either through dedicated interfaces online or through a download and subsequent analysis or editing offline. Most corpora offer an online interface with basic search functionality like full text search within the transcription and a simple visualization like Key Word in Context (KWIC). While this is enough for qualitative research in a uniform corpus, more complex tasks will become impossible if the data cannot be downloaded at all or only in a format that cannot be used by a specific user. We will motivate the necessity of these four kinds of access by example of common and concrete usage cases in the remainder of this section.

### 5.1 Search

Especially in spoken corpora with multiple speakers, one might want to restrict the results using the compiled metadata, e.g. by combining any query with the condition of only taking into account female speakers under the age of 35. This effectively enables the ad-hoc creation of sub-corpora (see Wynne 2008 for details) even if they were completely irrelevant for the original research question without any additional effort from the corpus builder or the user.

Furthermore, in a corpus with many different annotations and possibly multiple primary transcriptions, it is important to be able to search on specific layers in order to restrict results. This enables the user to search e.g. for *gehst du* without also finding

*gehste* which has been normalized to *gehst du* on another layer. This will also allow

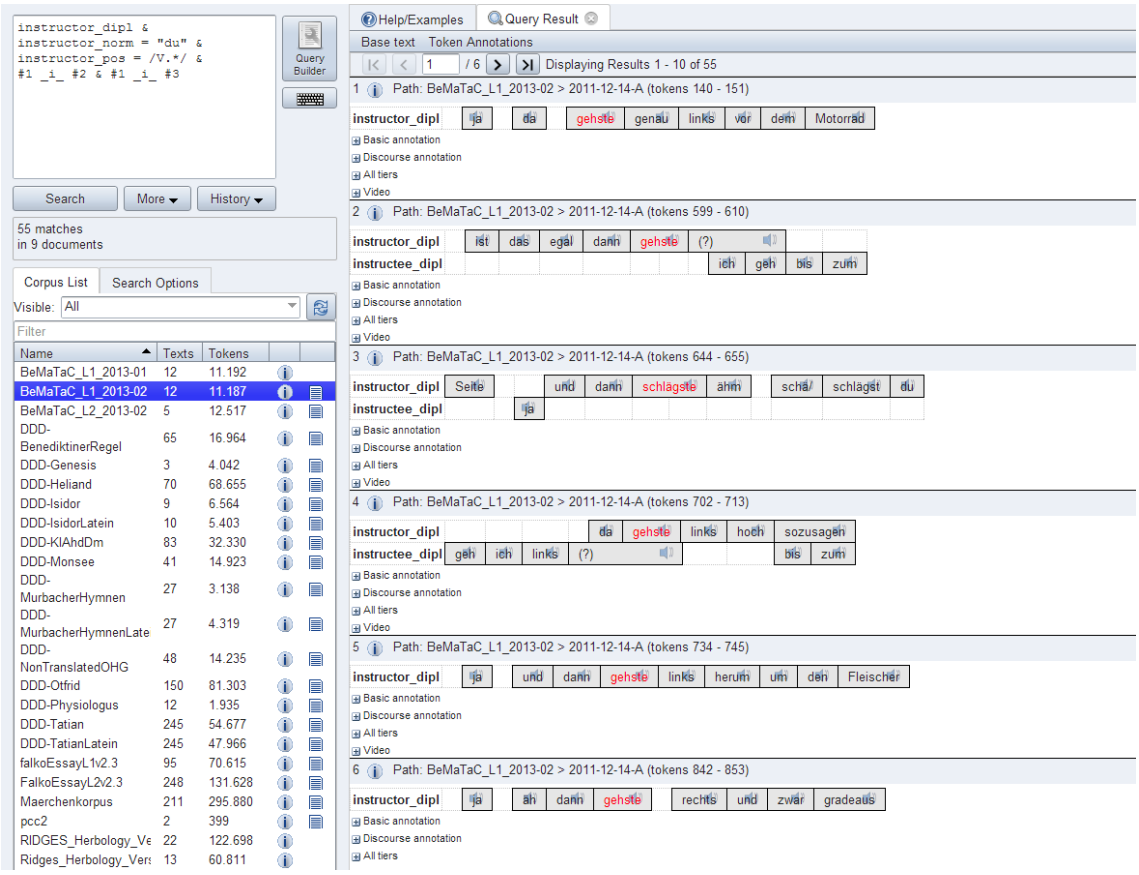


*Figure 7: BeMaTaC query and results in ANNIS, finding verbs with cliticized 2nd person singular pronoun 'du'.*

more complex and combining queries, such as searching for a specific word immediately followed by a word annotated with a specific part-of-speech tag. If one wanted to find out whether the normalized form *gehst du* is realized as *gehste*, one would have to write a complex query that compares the normalized transcription level with the narrower transcription level. If one wanted to find out which verbs occur with a cliticized 2nd person singular pronoun, one could formulate a query which uses the part-of-speech level referring to the normalized transcription and the narrow transcription, as shown in Figure 7.

In order to provide this kind of flexibility and also not having to redevelop the interface if the corpus is modified or extended, a formalized query language built on well-established concepts like logical operators and regular expressions can be used. BeMaTaC is primarily accessed via ANNIS (Zeldes et al. 2009, http://annis-tools.org), which uses AQL, a powerful node-and-edge-based query language. It is important to also publish the corpus' documentation and guidelines as to provide the user with information on which layers are available and what values they can contain.

### 5.2 Visualization

Different types of annotation need different types of visualizations, so users can better understand the data and see patterns; e.g. token-based annotations and span-based annotations can be visualized in tables, different types of syntactic trees need their own visualizations, and colors can help to see co-referent elements. If visualizations are developed in a general, annotation-agnostic way, they can be re-used for other purposes compatible with the same annotation format. ANNIS follows this idea by employing a modular approach: visualizations are rendered by pluggable Java modules, which can be developed and configured individually. In combination with a universal query language, the access interface as a whole can be used for a variety of corpora, removing the need to develop a separate interface for each new corpus. ANNIS simultaneously hosts a variety of written, spoken, historical, and learner corpora and even permits queries across multiple corpora.

As discussed in previous sections, any kind of transcription inevitably results in a loss of information. In order to make the transcription process transparent and to give access to the actual data that is ultimately described by the analyses conducted, aligned audio and/or video streams can be included in visualizations. In the case of BeMaTaC in

ANNIS, clicking on any token in a transcription or on any token or span annotation will automatically play back the corresponding segment of speech, as can be seen in Figure 8.

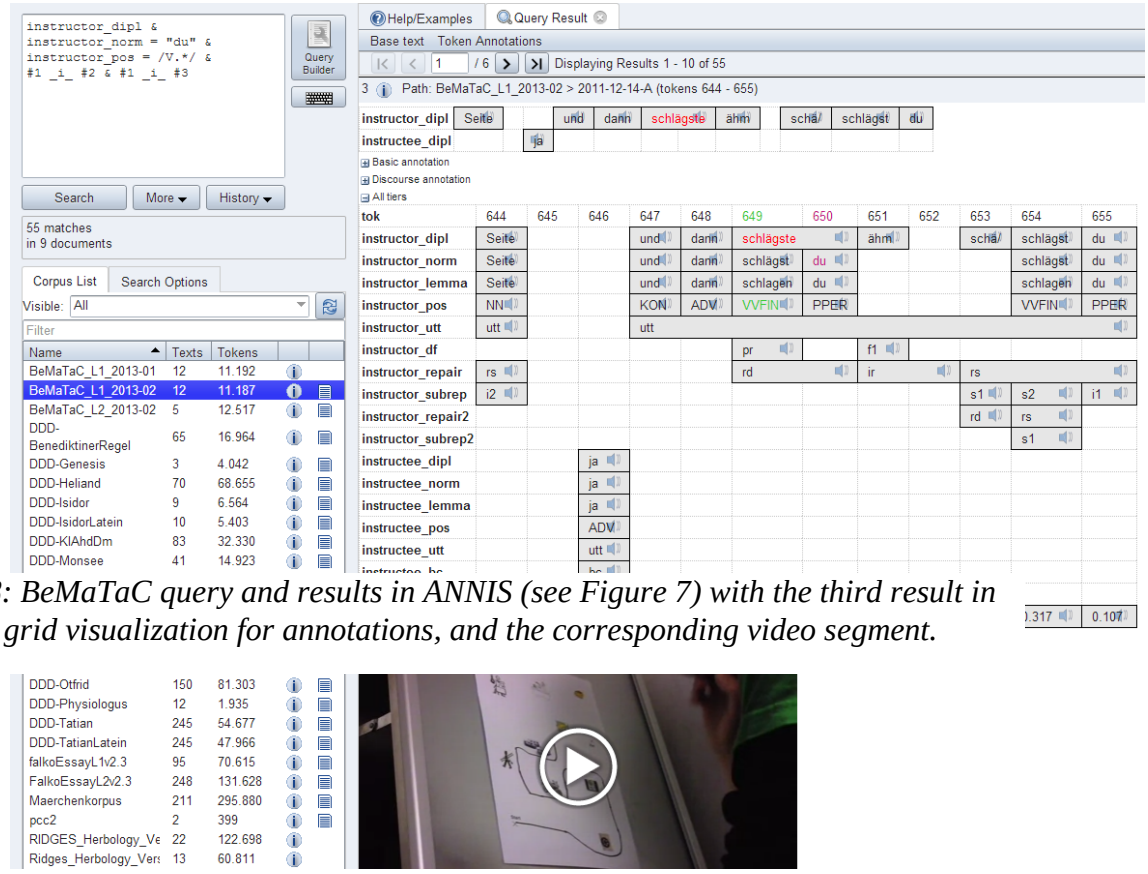


*Figure 8: BeMaTaC query and results in ANNIS (see Figure 7) with the third result in detail, a grid visualization for annotations, and the corresponding video segment.*

## 5.3 Download and export

For many questions it is sufficient to look at and analyze query results in a web interface. For some questions, however, it is necessary to download the original data (including some or all of the annotations) to perform special analyses or to add annotations. Providing a possibility to download a corpus in a variety of preferably standardized formats will enable users to apply their own software or scripts as well as common specialized software such as R (R Core Team 2013) or WEKA (Hall et al. 2009) for statistical analyses. For the reasons mentioned above, the original – and uncompressed – recordings should be made available. As a result, e.g. for research in

acoustic phonetics, spoken corpus data can become a valuable resource, even without any transcription.

Providing a search and visualization interface with the functionality to export data will result in an even more powerful tool, as it is possible to download only the data relevant for further analysis – such as an ad-hoc sub-corpus and/or the results of a query – and directly import it into the desired tool(s). This is also useful in order to further annotate data – by hand or automatically – if needed for a more specialized research question. Ideally, additional or modified data like this can even be fed back into the original corpus, allowing users to actively extend and improve the corpus. In BeMaTaC, a variety of annotations such as disfluencies, repairs, and backchanneling was not originally part of the corpus – researchers provided these along with their guidelines and documentation.

This type of information is crucial for the entire corpus and therefore has to be included in every release of the corpus. The exact circumstances of the recordings and their technical details need to be documented; speakers have to consistently provide metadata on a standardized form. The entire workflow of the corpus building process needs to be laid down; this includes extensive transcription and annotation guidelines as well as tools and their exact configurations. Without this documentation, the steps taken cannot be reproduced; findings therefore cannot be verified by the scientific community. In addition, future attempts at extending the corpus may fail if people originally involved in the process are no longer available.

## 6. Summary and conclusion

The need for transparency and reproducibility is basic in corpus linguistics (Lüdeling 2011). In this paper we argue that it is crucial for spoken corpora to be freely available

and that every step of interpretation – be it in transcription, in pre-processing such as tokenization, or in annotation – should be available with the auditory data. We show that this is possible if multi-layer standoff architectures are used. We present a specific multi-layer architecture that allows multiple and conflicting tokenizations, multiple speakers, multiple transcriptions per speaker, and completely speaker-independent data such as pauses as well as annotations in many different formats (Zeldes et al. 2009, Krause & Zeldes, 2014). Our main arguments for the need of such a flexible architecture are conceptual ((a) – (c)) as well as technical ((d)).

(a) Every transcription is an interpretation of the auditory data that is suitable to some research questions but excludes others. It should therefore be possible to add other transcriptions.

(b) Every pre-processing step is an interpretation. Even on the same transcription there could be different tokenizations. This could lead to differences in comparative statistics as well as differences in qualitative analyses. It should therefore be possible to see how a corpus is tokenized and to add different tokenizations.

(c) Every annotation is an interpretation. While this is true for ‘standard’ annotation layers such as part-of-speech layers it is even more important for annotation layers which are as yet less well understood such as disfluency layers. It should therefore be possible to see how categorization is done and to add further annotation layers wherever necessary.

(d) The conceptual need for a flexible architecture leads to considerable problems regarding formats and tool interoperability. This can be solved with a very

abstract general model and a converter framework such as SaltNPepper (Zipser & Romary 2010).

Providing corpus data in standoff architectures is not only necessary to make analyses transparent and reproducible, it also enables users to modify and extend the corpus according to their own needs and research questions, again independent of the nature of existing data or existing research questions. In our opinion, such a standoff architecture should be an integral part for state-of-the-art corpus distributions.

**Notes**

1. BeMaTaC was originally developed to provide a native-speaker reference corpus for the learner-exclusive Hamburg Map Task Corpus (HAMATAC, Schmidt et al. 2010) and therefore uses the same maps (Brinckmann et al. 2008) and basic design. However, in the course of working on BeMaTaC, we found a number of issues which we wanted to handle differently, which means that BeMaTaC, as it stands today, is not strictly comparable to HAMATAC.
2. The term ‘token’ is often used to mean something like a graphemic word but is technically defined as the smallest unit in a corpus, independent of what this smallest unit might be linguistically, see e.g. Schmid (2008).
3. In other corpus models the annotation is stored ‘inline’, i.e. in the same file as the primary data. Inline models are useful for large corpora with flat annotations because they can be searched very fast. They cannot be easily extended and are less flexible with respect to different annotation formats or conflicting annotations.
4. The Stuttgart-Tübingen-TagSet is the de facto standard for written German. It does not cover spoken phenomena such as disfluencies or cliticizations. For this reason, a

task force (led by Ulrich Heid and Heike Zinsmeister) is currently working on a revision.

5. Note, however, that there are currently two TEI special interest groups dealing with these questions – the SIG 'TEI for linguists' that is looking into extending the TEI to standoff models (Stührenberg 2012) and the SIG 'Computer-Mediated Communication' that tackles some issues like overlapping turns and non-standard forms that are also relevant for spoken language.
6. Salt cannot serve as a linguistic standard, as it is too abstract and does not constrain the data in any way.
7. This is true for all types of research data, as stated e.g. in the Berlin Declaration (Max-Planck Society 2014).

**References**

Anderson, A. H., M. Bader, E. Gurman Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, & R. Weinert. 1991. "The HCRC Map Task Corpus". *Language and Speech*, 34, 351-366.

Belz, M. 2013: online. *Disfluencies und Reparaturen bei Muttersprachlern und Lernern – eine kontrastive Analyse*. Humboldt-Universität zu Berlin. Available at: http://edoc.hu-berlin.de/docviews/abstract.php?id=40482 (accessed March 2014).

BeMaTaC. 2014: online. *BeMaTaC – A deeply annotated multimodal map-task corpus of spoken learner and native German*. Available at: http://u.hu-berlin.de/bematac (accessed March 2014).

Boersma, P. 2010. "Praat, a system for doing phonetics by computer". *Glot International*, 5 (9/10), 341-345.

Brinckmann, C., S. Kleiner, R. Knöbl, & N. Berend. 2008. “German Today: an areally extensive corpus of spoken Standard German”. In N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Paris: ELRA, 3185-3191.

Buchholz, S. & E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In L. Màrquez & D. Klein (Eds.), *Proceedings of the 10th Conference on Computational Natural Language Learning*. Stroudsburg: Association for Computational Linguistics, 149-164.

Burnard, L. (Ed.). 2007: online. *Reference Guide for the British National Corpus (XML Edition)*. Oxford: Research Technologies Service. Available at: http://www.natcorp.ox.ac.uk/XMLedition/URG (accessed March 2014).

Carletta J., S. Evert, U. Heid, J. Kilgour, J. Robertson, & H. Voormann. 2003. “The NITE XML Toolkit: flexible annotation for multi-modal language data”. *Behavior Research Methods, Instruments, & Computers*, 35 (3), 353-363.

Carletta J., S. Evert, U. Heid, & J. Kilgour. 2005. “The NITE XML Toolkit: data model and query”. *Language Resources and Evaluation*, 39 (4), 313-334.

Chiarcos, Ch., S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, & M. Stede. 2009. “A flexible framework for integrating annotations from different tools and tagsets”. *Traitement Automatique des Langues*, 49 (2), 271-291.

Creative Commons. 2014: online. *About The Licenses - Creative Commons*. Available at: http://creativecommons.org/licenses (accessed March 2014).

Dipper, S. 2005. “XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation”. In R. Eckstein & R. Tolksdorf (Eds.),

*Proceedings of Berliner XML Tage 2005*. Berlin: Humboldt-Universität zu Berlin, 39-50.

Dipper, S., A. Lüdeling, & M. Reznicek. 2013. “NoSta-D: A Corpus of German Non-Standard Varieties”. In M. Zampieri & S. Diwersy (Eds.), *Non-Standard Data Sources in Corpus-Based Research*. Aachen: Shaker, 69-76.

Druskat, S., L. Bierkandt, V. Gast, C. Rzymski, F. Zipser. 2014. “Atomic: an open-source software platform for multi-level corpus annotation”. In J. Ruppenhofer & G. Faaß (Eds.), *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014)*, 228–234. Available at: http://nbn-resolving.de/urn:nbn:de:gbv:hil2-opus-2866 (accessed May 2015).

Gerdes, K. 2014: online. *Arborator*. Available at: http://arborator.ilpga.fr (accessed March 2014).

Giesel, L., M. Klapi, D. Krüger, I. Nunberger, O. Rasskazova, & S. Sauer. 2013: online. “Berlin Map Task Corpus – A deeply annotated multimodal map-task corpus of spoken learner and native German”. *35. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. Available at: http://korpling.german.hu-berlin.de/bematac/publications/Giesel-et-al_2013_DGfS-CL-2013.pdf (accessed March 2014).

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, & I. H. Witten. 2009. “The WEKA Data Mining Software: An Update”. In O. R. Zaiane (Ed), *SIGKDD Explorations*, 11 (1), 10-18.

Hanke, T. & J. Storz. 2008. “iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography”. In O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd, & I. Zwitserlood (Eds.), *LREC 2008*

*Workshop, Proceedings, W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. Paris: ELRA, 64-67.

Himmelmann, N. P. 2012. “Linguistic Data Types and the Interface between Language Documentation and Description”. *Language Documentation & Conservation*, 6, 187-207.

Hinrichs, E. W., M. Hinrichs, & T. Zastrow. 2010. “WebLicht: Web-Based LRT Services for German”. In *ACL 2010 System Demonstrations, Proceedings*. Stroudsburg: Association for Computational Linguistics, 25–29.

Ide, N. & K. Suderman. 2007. GrAF: A Graph-based Format for Linguistic Annotations. In B. Boguraev, N. Ide, A. Meyers, Sh. Nariyama, M. Stede, J. Wiebe, & G. Wilcock (Eds.), *ACL 2007 Workshop, Proceedings, Linguistic Annotation Workshop*. Stroudsburg: Association for Computational Linguistics, 1–8.

Kirk, J. M. This volume. Solving the Spoken Language Paradox: The Pragmatically Annotated SPICE-Ireland Corpus.

Krause, T., A. Lüdeling, C. Odebrecht, & A. Zeldes. 2012: online. “Multiple Tokenization in a Diachronic Corpus”. *Exploring Ancient Languages through Corpora Conference 2012*. Available at: http://www.hf.uio.no/ifikk/english/research/projects/proiel/ealc/abstracts/Krause_et_al.pdf (accessed March 2014).

Krause, T. & A. Zeldes. 2014. “ANNIS3: A New Architecture for Generic Corpus Query and Visualization”. *Digital Scholarship in the Humanities.* Available at:

http://dsh.oxfordjournals.org/content/early/2014/12/02/llc.fqu057.full (accessed May 2015)

Lüdeling, A. 2011. “Corpora in Linguistics: Sampling and Annotation”. In K. Grandin (Ed.), *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. Stockholm: Center for History of Science at the Royal Swedish Academy of Sciences, 220-243.

Max Planck Society. 2014: online. *Max Planck Open Access: Berlin Declaration*. Available at: http://openaccess.mpg.de/Berlin-Declaration (accessed March 2014).

Müller, C. & M. Strube. 2006. “Multi-level annotation of linguistic data with MMAX2”. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy*. Frankfurt: Peter Lang, 197-214.

Nivre, J. 2008. “Treebanks”. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 225-241.

Pajas P. & J. Stepanek. 2008. “Recent Advances in a Feature-Rich Framework for Treebank Annotation”. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 673-680.

R Core Team. 2013: online. *R: A Language and Environment for Statistical Computing*. Available at: http://www.R-project.org (accessed March 2014).

Sauer, S. & O. Rasskazova. 2014: online. “BeMaTaC – eine digitale multimodale Ressource für Sprach- und Dialogforschung”. *Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen*. Available at: http://korpling.german.hu-

berlin.de/bematac/publications/Sauer-Rasskazova_2014_3WS-DHB.pdf (accessed March 2014).

Schiel, F., Ch. Draxler, & J. Harrington. 2011. “Phonemic Segmentation and Labelling using the MAUS Technique”. *Workshop New Tools and Methods for Very-Large-Scale Phonetics Research*, 28-31.

Schiller, A., S. Teufel, Ch. Stöckert, & Ch. Thielen. 1999: online. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* Available at: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (accessed March 2014).

Schmid, H. 1994: online. “Probabilistic Part-of-Speech Tagging Using Decision Trees”. In *Proceedings of International Conference on New Methods in Language Processing*. Available at: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf (accessed November 2014).

Schmid, H. 2008. “Tokenizing and Part-of-Speech Tagging”. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 527-551.

Schmidt, T. 2004. “Transcribing and annotating spoken language with EXMARaLDA”. In A. Witt, U. Heid, H. S. Thompson, J. Carletta, & P. Wittenburg (Eds.), *LREC 2004 Workshop, Proceedings, XML-based Richly Annotated Corpora*. Paris: ELRA, 69-74.

Schmidt, T. & K. Wörner. 2009. “EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research”. *Pragmatics*, 19 (4), 565-582.

Schmidt, T., H. Hedeland, T. Lehmberg, & K. Wörner. 2010: online. *HAMATAC – The Hamburg MapTask Corpus*. Available at: http://www.exmaralda.org/files/HAMATAC.pdf (accessed March 2014).

Sloetjes, H. & P. Wittenburg. 2008. "Annotation by category – ELAN and ISO DCR". In N. Calzolari, Kh. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Paris: ELRA, 816-820.

Stede, M. 2011. *Discourse Processing*. Morgan & Claypool.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, & J. Tsujii. 2012. "brat: a Web-based Tool for NLP-Assisted Text Annotation". In F. Segond (Ed.), *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics, 102-107.

Stührenberg, M. 2012: online. "The TEI and Current Standards for Structuring Linguistic Data". In P. Bański, E. Litta Modignani Picozzi, & A. Witt (Eds.), *Journal of the Text Encoding Initiative*, 3. Available at: http://jtei.revues.org/523 (accessed March 2014).

TEI Consortium. 2014: online. *TEI: Text Encoding Initiative*. Available at: http://www.tei-c.org (accessed March 2014).

Thompson, P. 2005: online. "Spoken Language Corpora". In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books, 59-70. Available at: http://ahds.ac.uk/linguistic-corpora (accessed March 2014).

Wichmann, A. 2008. “Speech corpora and spoken corpora”. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 187-207.

Wörner, K. 2009. *Werkzeuge zur flachen Annotation von Transkriptionen gesprochener Sprache*. Bielefeld: Bielefeld University.

Wynne, M. 2008. “Searching and Concordancing”. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter. 706-737.

Yimam, S. M., I. Gurevych, R. Eckart de Castilho, & C. Biemann. 2013. “WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations”. In M. Butt & S. Hussain (Eds.), *51st Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference System Demonstrations*. Stroudsburg: Association for Computational Linguistics, 1-6.

Zeldes, A, J. Ritz, A. Lüdeling, & Ch. Chiarcos. 2009: online. “ANNIS: A Search Tool for Multi-Layer Annotated Corpora”. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of Corpus Linguistics 2009*. Available at: http://edoc.hu-berlin.de/docviews/abstract.php?id=36996 (accessed March 2014).

Zipser, F. & L. Romary. 2010: online. “A model oriented approach to the mapping of annotation formats using standards”. In G. Budin, L. Romary, T. Declerck, P. Wittenburg (Eds.), *LREC 2010 Workshop, Proceedings, W4: Language Resource and Language Technology Standards*. Paris: ELRA. Available at: http://hal.inria.fr/inria-00527799 (accessed November 2014).