FOLKER: An Annotation Tool For Efficient Transcription Of Natural, Multi-Party Interaction

Thomas Schmidt, Wilfried Schütte
SFB 538 'Multilingualism'
Max Brauer-Allee 60
D-22765 Hamburg
E-mail: thomas.schmidt@uni-hamburg.de, schuette@ids-mannheim.de

Abstract

This paper presents FOLKER, an annotation tool developed for the efficient transcription of natural, multi-party interaction in a conversation analysis framework. FOLKER is being developed at the Institute for German Language in and for the FOLK project, whose aim is the construction of a large corpus of spoken present-day German, to be used for research and teaching purposes. FOLKER builds on the experience gained with multi-purpose annotation tools like ELAN and EXMARaLDA, but attempts to improve transcription efficiency by restricting and optimizing both data model and tool functionality to a single, well-defined purpose. This paper starts with a description of the GAT transcription conventions and the data model underlying the tool. It then gives an overview of the tool functionality and compares this functionality to that of other widely used tools.

1. Introduction

This paper presents FOLKER, an annotation tool developed for the efficient transcription of natural, multi-party interaction in a conversation analysis framework. FOLKER is being developed at the Institute for German Language in and for the FOLK project, whose aim is the construction of a large corpus of spoken present-day German, to be used for research and teaching purposes (see FOLK 2010). FOLKER builds on the experience gained with multi-purpose annotation tools like ELAN and EXMARaLDA (Rohlfing et al. 2006), but attempts to improve transcription efficiency by restricting and optimizing both data model and tool functionality to a single, well-defined purpose.

2. FOLK Corpus

FOLK (Forschungs- und Lehrkorpus) is the "Research and Teaching Corpus of Spoken German". Recognizing that there is, to date, no larger, systematically stratified collection of publicly available recordings of authentic spoken interaction, let alone a consistent set of corresponding, computer-accessible transcriptions for German, the Pragmatics Department of the Institute for German Language (IDS) started to set up FOLK in 2008. Recordings for the corpus are partly collected from other sources (the institute's spoken language archive and other corpora of talk in interaction collected outside the IDS), partly done from scratch for the project. The aim is to cover a broad spectrum both in terms of regional variation and in terms of different interaction types. All recordings are transcribed within the project. In order to ensure a high level of consistency, an efficient transcription workflow, high community acceptance and good automatic processability of the data, a group of conversation analysis researchers and a group of software developers were actively involved in the planning stage of the corpus. By coordinating corpus development, the specification of transcription conventions (see next section) and the development of an annotation tool (i.e. FOLKER) in this way, we also hope to contribute to the establishment of a best practice in our field.

3. GAT Transcription Conventions

The GAT transcription conventions (see Selting et al. 1998 and 2009) are a de-facto standard in Germany for the transcription of natural interaction in conversation analytic research. Since, however, they originally disregarded the question of an adequate computer encoding, an initiative was started to revise the conventions and modify them such that GAT transcriptions could be represented in a formal data model and a corresponding XML file format (see Schmidt 2007 and Schmidt et al. 2008, also next section).

```
0001 PRE good evening
0002 AUD ((laughter))
0003 PRE i have with me (.) tonight (0.3)
       ann elk
0004     mistress ann elk
0005 ELK (0.2) miss
0006 PRE (0.7) you
0007     °hh have a new theory the
       brontosaurus
0008 ELK well ehm can i just eh say here
       chris for one moment
0009     that i have a new theory about the
       brontosaurus
0010 AUD ((laughter))
0011 PRE exactly
0012 AUD ((laughter))
0013 PRE what is it
```

Figure 1: GAT transcript

The revised version of GAT (Selting et al. 2009) now defines three transcription levels which correspond to different degrees of prosodic detail. FOLK and FOLKER only make use of the first of these levels – the minimal

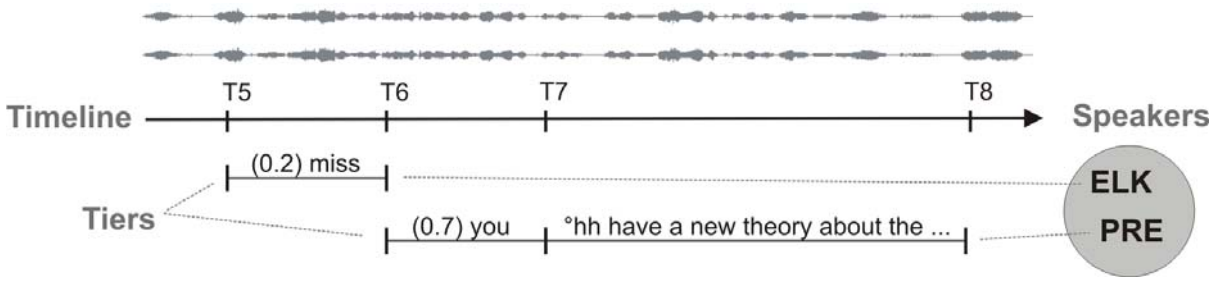

Figure 2: Single timeline, multiple tiers (STMT) data model

transcript – which provides rules for the transcriptions of words in a modified orthography, for the description of pauses, breathing and non-phonological phenomena (coughing, laughing etc.) and for the handling of uncertain or incomprehensible passages. Figure 1 gives an example of a GAT minimal transcription as a plain text file.[1]

4. Data Model / Data Format

FOLKER's basic data model is derived from the single-timeline-multiple-tiers (STMT) data model described in (Schmidt 2005), i.e. it conceptualizes a transcription as a set of annotations which are assigned via a start and an end point to a single, fully ordered timeline, and which are partitioned into a number of tiers such that no two annotations within a tier overlap.

```
<folker-transcription>
  <speakers>
    <speaker id="PRE"/>
    <speaker id="ELK"/>
    <speaker id="AUD"/>
  </speakers>
  <recording path="./MyTheory.wav"/>
  <timeline>
    <!-- [...] -->
    <timepoint id="T5" time="18.79"/>
    <timepoint id="T6" time="20.18"/>
    <timepoint id="T7" time="23.08"/>
    <timepoint id="T8" time="25.99"/>
    <!-- [...] -->
  </timeline>

<!-- [...] -->
  <contribution speaker="ELK" start="T5" end="T6">
    <pause duration="0.2"/>
    <w>miss</w>
  </contribution>
  <contribution speaker="PRE" start="T6" end="T8">
    <pause duration="0.7"/>
    <w>you</w>
    <time reference="T7"/>
    <breathe type="in" length="2"/>
    <w>have</w>
    <w>a</w>
    <w>new</w>
    <w>theory</w>
    <w>about</w>
    <w>the</w>
    <w>brontosaurus</w>
  </contribution>
</folker-transcription>
```

Figure 3: Folker data format

[1] We use an English example here for illustration purposes. In the FOLK project, FOLKER is of course used exclusively for transcriptions of German.

As a further specification and restriction, the FOLKER data model requires that each tier be assigned to a speaker, and that no two tiers can be assigned to the same speaker. Figure 2 illustrates this for an excerpt from the above example.

While this data model represents the temporal structure of events, including possible overlaps between different speakers, further structure is added by combining adjacent annotations into contributions and re-segmenting them into the entities defined in the transcription convention (i.e. words, pauses etc.) with the help of a finite state transducer (the process is described in more detail in Schmidt 2005). The resulting data structure can be serialized into a TEI-like XML format as illustrated in figure 3. Since the temporal information in the data model is structurally compatible with the data models of tools like ELAN or EXMARaLDA, FOLKER can provide export filters for these tools.

5. Tool functionality

FOLKER's main interface offers three editable views of the transcription data. Each of these views is optimized for a specific step in the transcription workflow, and users can freely switch between the views at any time in the transcription process.

5.1 Segment view

The segment view, illustrated in figure 4, is most efficient for initial transcription. It displays individual annotations in a vertical list, thus optimally exploiting screen real estate and giving the transcriber a more text-like feeling of the transcription than horizontally organized display methods (like musical scores) do. Speaker assignment, annotation text and temporal assignment can be freely modified in this view and individually for each annotation.

Using a regular expression, the syntax of the annotation text is checked during input for conformance with the transcription conventions. Errors (as the missing closing bracket in line 10 in figure 4) are indicated by a red X in the column labelled 'Syntax'. Likewise, the temporal integrity of annotations is verified. If two annotations belonging to the same speaker overlap (which the data model prohibits), this is indicated in the column labelled 'Zeit'.

5.2 Partitur view

The Partitur (musical score) view, illustrated in figure 5, displays the same transcription in a horizontal layout,

organised into tiers. This view, which is comparable to the main interfaces of tools like ANVIL, ELAN, EXMARaLDA or Praat, is best suited for editing temporal relations, most prominently speaker overlap. Important operations in this view include splitting and merging annotations and shifting characters between annotations.

5.3 Contribution view

The contribution view, finally, also uses a vertically organised layout, but, instead of individual annotations, displays adjacent annotations of speakers as contributions. This view is thus close to a traditional, drama-script like representation of an interaction, complying with established reading habits. It is therefore best suited for final proof-reading and corrections of a transcript. As in the segment view, additional columns give information about the syntactic correctness and temporal integrity of the transcribed data.

5.4 Audio alignment

Navigation in all three views is synchronized with navigation in (a waveform visualization of) the audio recording. The audio player provides buttons for playing or looping the current selection, and for playing the last second of the current selection, the latter functionality being important for an efficient fine tuning of segment boundaries. This fine tuning can be carried out either by dragging segment boundaries with the mouse, by scrolling the mouse wheel up or down in the vicinity of a boundary, or by using keyboard shortcuts. Since segments refer to an explicit timeline, rather than directly to times in the recording, a modification of a segment boundary often also affects the boundaries of neighbouring segments. This makes it easier for the transcriber to keep the temporal integrity of annotations intact.

5.5 Other functionality

Additional functionality of FOLKER includes:

- Support for multi-part transcriptions: this is important for very long recordings (> 90 minutes) whose transcription has to be distributed over several documents in order to ensure a reasonable processing performance;
- Export routines for EXMARaLDA and ELAN data and an import routine for EXMARaLDA data;
- Visualisation functions for displaying a transcription as a segment list, a musical score or a contribution list in a browser or a text processor;
- Search and search&replace routines;
- Automatic procedures for filling gaps in the transcription, for normalizing whitespace and for measuring pauses.

6. Comparison with other tools

In contrast to many other widely used transcription tools, FOLKER is explicitly and consciously designed not as a multi-purpose tool, but rather as a tool which supports one specific (albeit widely used) annotation scenario in a maximally efficient way. If we outline the tool's particular strengths in comparison with other tools in the remainder of this section, we therefore do not want this to be understood as a claim that FOLKER is superior to these tools in general. Quite to the contrary, we believe that researchers should be encouraged to exploit the growing interoperability between different solutions and use tools with different specializations for different tasks in their corpus construction workflows. In such a workflow, we see FOLKER as a tool for the creation of base transcriptions which can later be supplemented with additional annotations through other tools.

We restrict our comparison to tools of which we know that they have been used to construct conversation or discourse corpora.

6.1 FOLKER vs. Transana and similar tools

As far as their transcription functionality is concerned, tools like Transana[2] are basically combinations of a text editor with a media player, i.e. they offer the possibility to type a free, possibly formatted text (the transcription) and add some special functions for navigating in a recording and for linking pieces of text to that recording.

While the similarity of these tools to ordinary text processing software makes them popular among "naive" users, who value the resulting "ease of use", they do virtually nothing to ensure adequate computer processability of the data – since their data structure is that of a free text, the formats do not contain any structural information that could be systematically exploited by a database or some other advanced application. FOLKER is much more ambitious in this respect – it produces data in an open standard format which validates against a specific data schema in which all the relevant information is encoded in a methodical manner. It is thus suitable for reliable and systematic querying and for other types of (semi-)automatic processing.

6.2 FOLKER vs. CLAN

CLAN, the transcription and annotation software provided by the CHILDES[3] system is comparable to FOLKER insofar as it also closely tied to a specific transcription convention (CHAT) and also has functionality for checking the conformity of a transcribed text with respect to these conventions.

FOLKER differs from CLAN, first of all, in its more modern user interface. Second, FOLKER offers different editable views on the data whereas CLAN is restricted to a single, line-based view (comparable to FOLKER's contribution view).

Third, although CLAN is able to write an XML-based file format, it does not use an underlying data model in the strict sense of the word, i.e. in the sense of an algebraically well-defined formalism. FOLKER, in contrast, is based on the idea of annotation graphs (AGs), or, more specifically, on the STMT subset of AGs, thus making it much easier to validate and exploit certain structural features of the data.

[2] http://www.transana.org/
[3] http://childes.psy.cmu.edu/

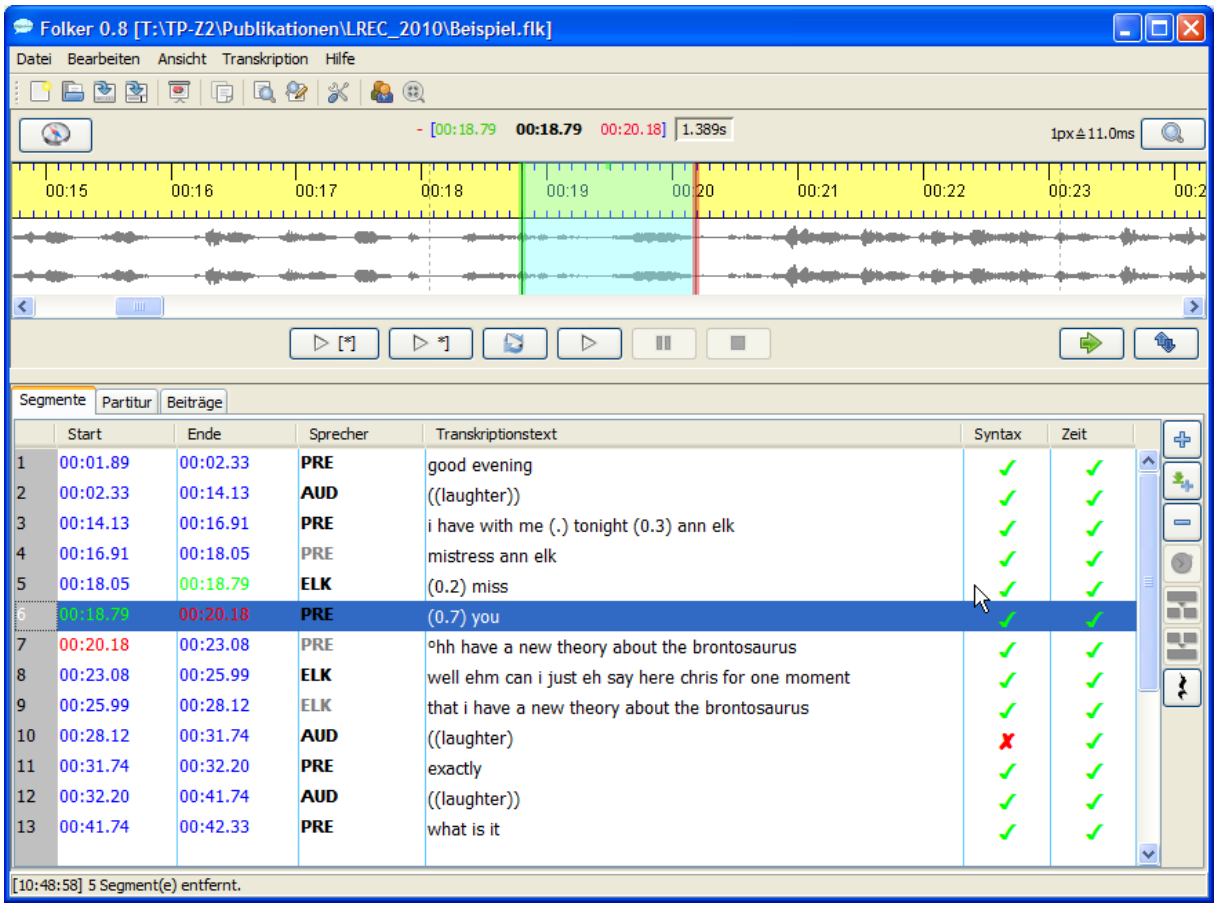

Figure 4: Segment view

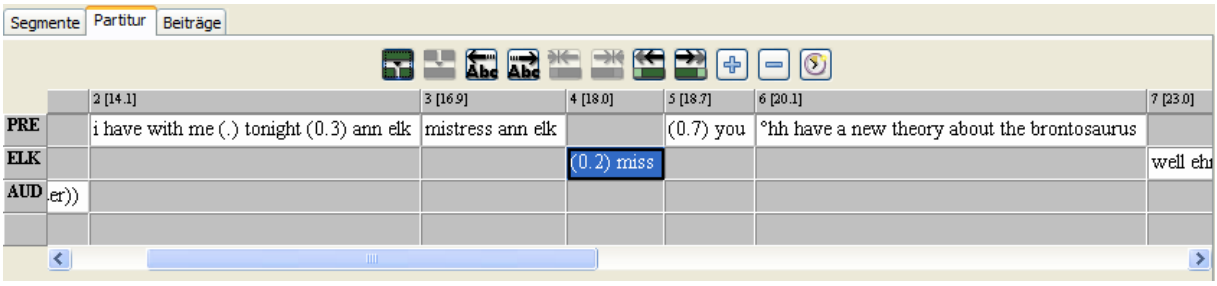

Figure 5: Partitur view

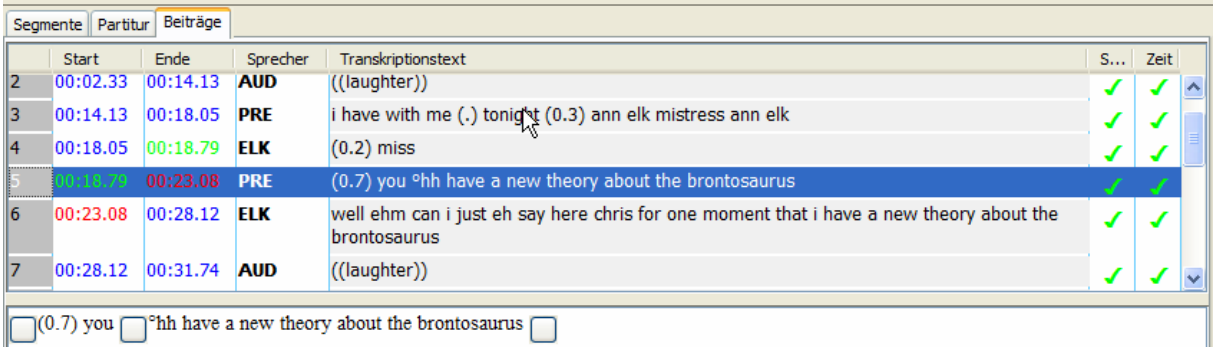

Figure 6: Contribution view

On the other hand, CLAN is of course applicable to a much wider range of annotation tasks. Most importantly, it provides a large number of so-called dependent tiers in which the main (transcription) tier can be supplemented with further analytic information. FOLKER does not cater for such multi-level annotation tasks.

6.3 FOLKER vs. Transcriber

Transcriber[4] is comparable to FOLKER insofar as it was also optimized to support a specific transcription and annotation task (broadcast speech) in a maximally efficient manner. Also similar to FOLKER, Transcriber offers different views of the data – a line-based view comparable to FOLKER's contribution view and a time-based view comparable to FOLKER's partitur view. However, unlike in FOLKER, only the first of these views is editable in Transcriber, the second mainly serving as a help for orientation in the whole document. Moreover, FOLKER has a more general approach to the handling of multiple speaker scenarios. Most importantly, Transcriber treats overlaps as separate structural entities which are assigned to multiple speakers. In FOLKER, on the other hand, speaker assignment is carried out independently of temporal speaker constellations, thus making the representation of more complex types of speaker overlaps (e.g. partial overlap between more than two speakers) easier and more flexible.

6.4 FOLKER vs. ELAN and similar tools

Although the tools share a common basis in their time-based data models, the functionality of ELAN[5] (and similar tools like, e.g., ANVIL) is generally much more comprehensive than that of FOLKER. To start with, ELAN supports different media types (audio and video) and formats, while FOLKER is restricted to WAV audio. Second (like CLAN and EXMARaLDA) ELAN is designed for multi-level annotation whereas FOLKER is restricted to one transcription layer per speaker. Third, ELAN contains a multitude of functionality that is not directly related to the transcription or annotation task, but addresses more far-reaching needs like metadata description or query. Fourth, unlike FOLKER, ELAN is not fixed on a single annotation scenario, transcription convention or coding scheme, but supports many such scenarios.

In sum, FOLKER is thus a considerably less powerful piece of software than ELAN, but this reduced complexity also makes it more accessible for many users with a lower level of technical know-how.

What furthermore distinguishes FOLKER from ELAN is its approach to the representation of (possibly competing, non-hierarchizable) temporal and linguistic structure of speaker's utterances. In ELAN, annotations in main tiers referring to the timeline can be segmented (e.g. into words or other tokens) on another tier using the concept of a so-called "symbolic subdivision". Different structural divisions are thus represented on different layers of the data model. In contrast, FOLKER (like EXMARaLDA) integrates such linguistic segmentations and the temporal structure of the discourse into one layer of the representation (see figure 3 and Schmidt 2005).

[4] http://trans.sourceforge.net/en/presentation.php

[5] http://www.lat-mpi.eu/tools/tools/elan

6.5 FOLKER vs. EXMARaLDA Partitur-Editor

Since FOLKER's data model is a subset of EXMARaLDA's[6] data model and the tools also share a fair proportion of their code base, there is a lot of commonality between them. The main difference between the tools is to be found at the user interface level. The EXMARaLDA Partitur-Editor offers a single view on the data which is more or less identical to FOLKER's partitur view. Like ELAN and in contrast to FOLKER, EXMARaLDA supports multi-level annotation of different media types and formats. Again, FOLKER is thus a less powerful, but also a more easily accessible tool. The close relation between the two tools, however, makes it very easy to use them side-by-side.

7. Availability / Outlook

FOLKER has now been tested for more than a year and is confirmed to run reliably on different Windows operating systems (XP, Vista and Windows 7). A Macintosh version is also available, but this has received less attention in the test phase.

For about a year, FOLKER has been used productively not only inside the FOLK project, but also in other spoken language corpus projects, for example at the Research Centre on Multilingualism for the construction of a corpus documenting language attrition in speakers of Italian.

At the current stage, no essential extension of the existing functionality is planned, but we will continue to improve and optimise the existing functionality.

A tool for orthographic normalization of FOLKER transcriptions (i.e. for annotation of modified orthographic forms in a GAT transcription with their standard lemmas) is under construction.

FOLKER is available freely for use in academic research and teaching. It can be downloaded after registration from the website of the IDS Pragmatics Department at http://agd.ids-mannheim.de/html/folker.shtml.

8. Acknowledgements

The work described in this paper has been financed by the Pragmatics Department of the Institute for the German Language (IDS) in Mannheim. FOLKER is built in parts on EXMARaLDA technology, developed at the Research Centre on Multilingualism (University of Hamburg) with a grant by the German Science Foundation (DFG).

9. References

FOLK (2010). Website of the FOLK corpus at the IDS Mannheim. http://agd.ids-mannheim.de/html/folk.shtml

Rohlfing, K.; Loehr, D.; Duncan, S.; Brown, A.; Franklin, A.; Kimbara, I.; Milde, J.; Parrill, F.; Rose, T.; Schmidt, T.; Sloetjes, H.; Thies, A. & Wellinghoff, S. (2006) Comparison of multimodal annotation tools — workshop report. In: *Gesprächsforschung* (7), pp. 99-123.

Schmidt, T. (2007): Transkriptionskonventionen für die computergestützte gesprächsanalytische Transkription. In: *Gesprächsforschung* (8), pp. 229-241. http://www.gespraechsforschung-ozs.de.

[6] http://www.exmaralda.org

Schmidt, T.; Deppermann, A.; Hartung, M.; Schütte, W. (2008) GAT: Aspekte der computertechnischen Umsetzbarkeit. Technical report Universität Hamburg / IDS Mannheim. http://www.exmaralda.org.

Schmidt, T. (2005): Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. In: *Working papers in multilingualism*, Series B (62).

Selting, M.; Auer, P.; Barden, B.; Bergmann, J.; Couper-Kuhlen, E.; Günthner, S.; Meier, C.; Quasthoff, U.; Schlobinski, P. & Uhmann, S. (1998) Gesprächsanalytisches Transkriptionssystem (GAT). In: *Linguistische Berichte* 173, pp. 91-122.

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