EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language

Thomas Schmidt
Institut für Deutsche Sprache, Mannheim
R 5, 6-13
D-68161 Mannheim
thomas.schmidt@uni-hamburg.de

Abstract

This paper presents two toolsets for transcribing and annotating spoken language: the EXMARaLDA system, developed at the University of Hamburg, and the FOLK tools, developed at the Institute for the German Language in Mannheim. Both systems are targeted at users interested in the analysis of spontaneous, multi-party discourse. Their main user community is situated in conversation analysis, pragmatics, sociolinguistics and related fields. The paper gives an overview of the individual tools of the two systems – the Partitur-Editor, a tool for multi-level annotation of audio or video recordings, the Corpus Manager, a tool for creating and administering corpus metadata, EXAKT, a query and analysis tool for spoken language corpora, FOLKER, a transcription editor optimized for speed and efficiency of transcription, and OrthoNormal, a tool for orthographical normalization of transcription data. It concludes with some thoughts about the integration of these tools into the larger tool landscape.

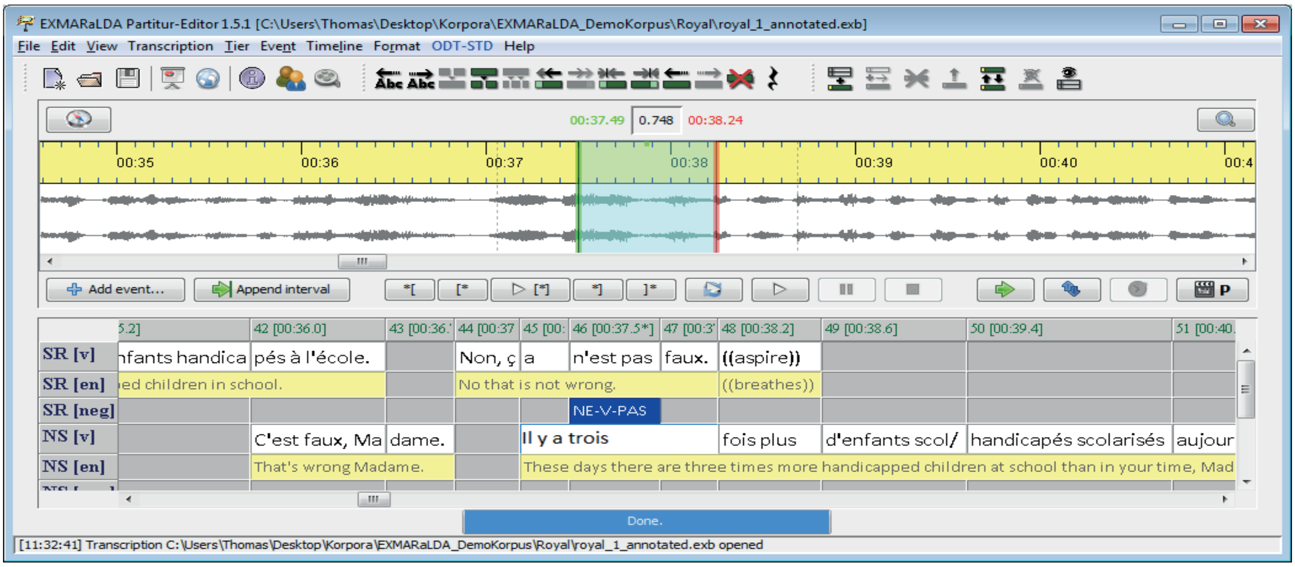
Keywords: transcription, annotation, software tools, corpus construction, corpus analysis

1. Introduction

The transcription and annotation of audio and video recordings is an important method not only in speech technology, but also in many fields of research in the humanities and the social sciences. Building and sharing bigger corpora of such transcriptions is a key challenge for all those fields because it is often only through sufficiently large amounts of data that meaningful results can be obtained. Computational tools for spoken language corpus building play a central role in that process because they determine not only the quality and efficiency with which the tedious manual procedure of transcription and annotation is carried out, but also the possible uses of the resulting data.

Since the research interests, the expectations towards computational support, and the concrete practices of transcription and annotation differ greatly across and within the disciplines, different computer tools have been developed which all take slightly different approaches to the same task. Thus, Praat focuses on the phonetic analysis of audio data, ELAN owes much of its design to the requirements of the language documentation community, Transcriber is optimized for the transcription of broadcast speech, CLAN for the transcription and annotation of child language, and so forth.

In this paper, I will present two tool sets whose background is the analysis of spontaneous, multi-party discourse as it is practiced in conversation analysis, pragmatics, sociolinguistics and related fields. I will give an overview of their functionality



EXMARaLDA Partitur-Editor 1.5.1 [C:\Users\Thomas\Desktop\Korpora\EXMARaLDA_DemoKorpus\Royal\royal_1_annotated.exb]

	5.2]	42 [00:36.0]	43 [00:36.'	44 [00:37	45 [00:	46 [00:37.5*]	47 [00:3'	48 [00:38.2]	49 [00:38.6]	50 [00:39.4]	51 [00:40.
SR [v]	ɲfants handica	pés à l'école.		Non, ç	a	n'est pas	faux.	((aspire))			
SR [en]	ed children in school.			No that is not wrong.				((breathes))			
SR [neg]						NE-V-PAS					
NS [v]		C'est faux, Ma	dame.		Il y a trois			fois plus	d'enfants scol/	handicapés scolarisés	aujour
NS [en]		That's wrong Madame.			These days there are three times more handicapped children at school than in your time, Mad						

[11:32:41] Transcription C:\Users\Thomas\Desktop\Korpora\EXMARaLDA_DemoKorpus\Royal\royal_1_annotated.exb opened

Figure 1: Partitur-Editor

and then discuss some aspects of their role in the larger tool landscape.

2. EXMARaLDA

EXMARaLDA (Extensible Markup Language for Discourse Annotation) was developed at the Research Centre on Multilingualism at the University of Hamburg between 2000 and 2011. It defines a data model for multi-layer annotation of media files which can be understood as a semantically specified and structurally simplified version of the more generic framework of annotation graphs (Bird/Liberman 2001). The data model is usually represented in XML file formats, but can also be mapped to an RDB representation. The creation, management and analysis of EXMARaLDA data is enabled through three principal tools which I will briefly introduce in the following sections.

2.1. EXMARaLDA Partitur-Editor

The Partitur-Editor is a tool for creating transcriptions of audio or video recordings in a musical score (German: Partitur) interface, i.e. in a two-dimensional layout in which time progresses from left to right and different annotation types are organized in tiers on the vertical axis.

During transcription or in a separate step, the transcribed text can be linked to the underlying audio or video file by setting appropriate timestamps in the transcription's timeline. It is possible to associate a transcription with more than one media file (e.g. a video and an audio recording or two video recordings from different perspectives) and to switch between them, provided that they are synchronized. Tiers and speakers can be freely added or modified at any time in the transcription process. Likewise, timestamps can be adjusted or fine-tuned efficiently after they have been set. In order to better distinguish different data types or speakers, each tier can be given an individual formatting (e.g. a different font type or a background coloring). The data model requires one tier per speaker to be defined as the main tier containing either an orthographic or a phonetic transcription. Any number of secondary annotation tiers can be added to the main tier containing, for example, prosodic annotations, utterance translations, etc.

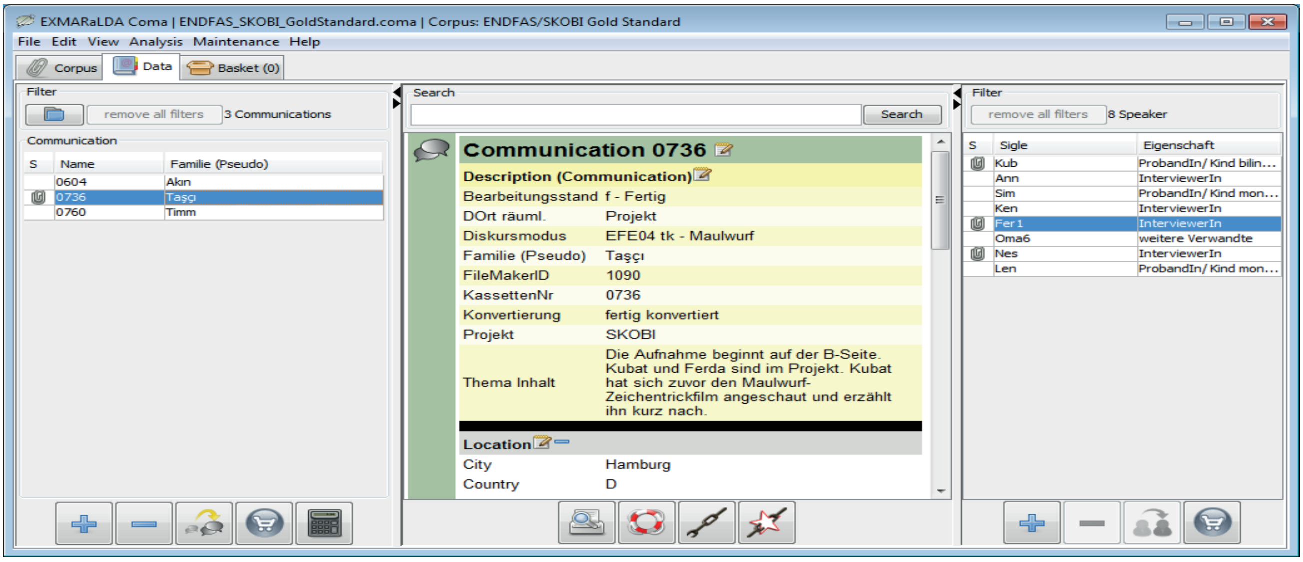
Further structural information can be added to the data through the process of *segmentation* which identifies linguistic entities in the temporally aligned annotations and thus makes them available for computerized analysis. Which entities are identified and how they are segmented depends on the transcription system used. For example, segmentation according to the HIAT transcription system will subdivide the orthographic transcriptions of speakers into utterances, which, in turn, decompose into words, pauses and descriptions of non-verbal behaviour (e.g. "coughs"). Other transcription systems may use intonation phrases instead of utterances, distinguish different subtypes of non-verbal behavior, etc. Currently, the most important German systems for conversation and discourse analysis (HIAT – Rehbein et al. 2004, GAT – Selting et al. 1998 and DIDA – Schütte 2004) are supported, as well as the CHAT system of CHILDES (MacWhinney 2000) and an IPA based system.

2.2. EXMARaLDA Corpus Manager

Many corpora in conversation analysis, in sociolinguistics and in related fields follow a design in which interactions and speakers are differentiated according to a great number of variables, such as age or educational status of speakers, interaction type, etc.

In order to adequately represent and manage the speaker and interaction metadata for such corpora, EXMARaLDA provides the Corpus Manager, a tool for assembling recordings and transcriptions into a corpus and describing the entities it is made up of through a number of appropriate metadata attributes. These metadata are then used both for cataloging and describing the resource for reuse (e.g. in a digital infrastructure like CLARIN) and for correlating linguistic and extra-linguistic facts during analysis.

The Corpus Manager not only supports the user in building and administering a spoken language corpus, but also provides functionality for corpus maintenance (e.g. checking the integrity of media links or the consistency of annotations), for filtering transcriptions according to metadata (e.g. only transcriptions of native speakers older than ten years), and for simple analyses of a corpus as a whole (e.g. generating a corpus statistics or a corpus token list).


EXMARaLDA Coma | ENDFAS_SKOBI_GoldStandard.coma | Corpus: ENDFAS/SKOBI Gold Standard
File Edit View Analysis Maintenance Help
Corpus
Data
Basket (0)
Filter
remove all filters
Search
Search
Filter
remove all filters

3 Communications

Communication

S	Name	Familie (Pseudo)
	0604	Akın
	0736	Taşçı
	0760	Timm

Communication 0736

Description (Communication)

Bearbeitungsstand	f - Fertig
DOrt räuml.	Projekt
Diskursmodus	EFE04 tk - Maulwurf
Familie (Pseudo)	Taşçı
FileMakerID	1090
KassettenNr	0736
Konvertierung	fertig konvertiert
Projekt	SKOBI
Thema Inhalt	Die Aufnahme beginnt auf der B-Seite. Kubat und Ferda sind im Projekt. Kubat hat sich zuvor den Maulwurf-Zeichentrickfilm angeschaut und erzählt ihn kurz nach.

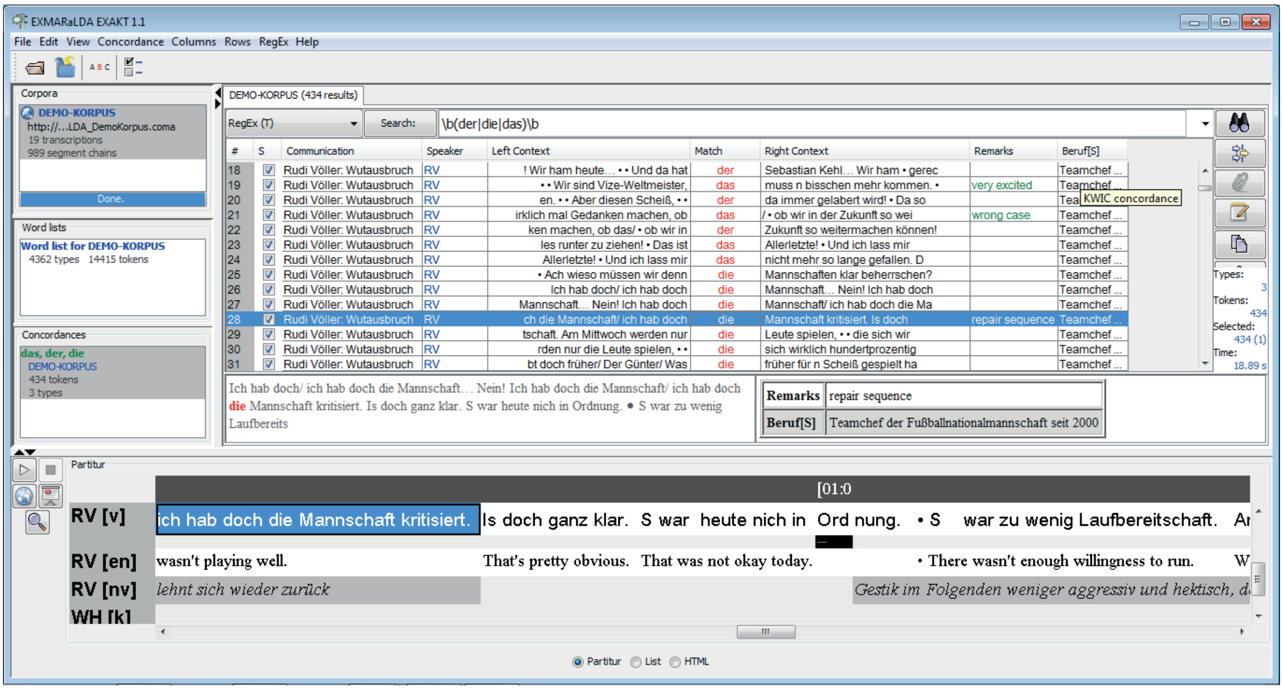
Location

City	Hamburg
Country	D

8 Speaker

S	Sigle	Eigenschaft
	Kub	ProbandIn/ Kind bilin...
	Ann	InterviewerIn
	Sim	ProbandIn/ Kind mon...
	Ken	InterviewerIn
	Fer1	InterviewerIn
	Oma6	weitere Verwandte
	Nes	InterviewerIn
	Len	ProbandIn/ Kind mon...

Figure 2: Corpus Manager



\b(der|die|das)\b

#	S	Communication	Speaker	Left Context	Match	Right Context	Remarks	Beruf[S]
18	☑	Rudi Völler: Wutausbruch	RV	! Wir ham heute… • • Und da hat	der	Sebastian Kehl… Wir ham • gerec		Teamchef ...
19	☑	Rudi Völler: Wutausbruch	RV	• • Wir sind Vize-Weltmeister,	das	muss n bisschen mehr kommen. •	very excited	Teamchef ...
20	☑	Rudi Völler: Wutausbruch	RV	en. • • Aber diesen Scheiß, • •	der	da immer gelabert wird! • Da so		Tea
21	☑	Rudi Völler: Wutausbruch	RV	irklich mal Gedanken machen, ob	das	/ • ob wir in der Zukunft so wei	wrong case	Teamchef ...
22	☑	Rudi Völler: Wutausbruch	RV	ken machen, ob das/ • ob wir in	der	Zukunft so weitermachen können!		Teamchef ...
23	☑	Rudi Völler: Wutausbruch	RV	les runter zu ziehen! • Das ist	das	Allerletzte! • Und ich lass mir		Teamchef ...
24	☑	Rudi Völler: Wutausbruch	RV	Allerletzte! • Und ich lass mir	das	nicht mehr so lange gefallen. D		Teamchef ...
25	☑	Rudi Völler: Wutausbruch	RV	• Ach wieso müssen wir denn	die	Mannschaften klar beherrschen?		Teamchef ...
26	☑	Rudi Völler: Wutausbruch	RV	Ich hab doch/ ich hab doch	die	Mannschaft… Nein! Ich hab doch		Teamchef ...
27	☑	Rudi Völler: Wutausbruch	RV	Mannschaft… Nein! Ich hab doch	die	Mannschaft/ ich hab doch die Ma		Teamchef ...
28	☑	Rudi Völler: Wutausbruch	RV	ch die Mannschaft/ ich hab doch	die	Mannschaft kritisiert. Is doch	repair sequence	Teamchef ...
29	☑	Rudi Völler: Wutausbruch	RV	tschaft. Am Mittwoch werden nur	die	Leute spielen, • • die sich wir		Teamchef ...
30	☑	Rudi Völler: Wutausbruch	RV	rden nur die Leute spielen, • •	die	sich wirklich hundertprozentig		Teamchef ...
31	☑	Rudi Völler: Wutausbruch	RV	bt doch früher/ Der Günter/ Was	die	früher für n Scheiß gespielt ha		Teamchef ...

Ich hab doch/ ich hab doch die Mannschaft… Nein! Ich hab doch die Mannschaft/ ich hab doch **die** Mannschaft kritisiert. Is doch ganz klar. S war heute nich in Ordnung. ● S war zu wenig Laufbereits

Remarks	repair sequence
Beruf[S]	Teamchef der Fußballnationalmannschaft seit 2000

	[01:0					
RV [v]	ich hab doch die Mannschaft kritisiert.	Is doch ganz klar.	S war heute nich in	Ord nung.	• S war zu wenig Laufbereitschaft.	Ar
RV [en]	wasn't playing well.	That's pretty obvious.	That was not okay today.		• There wasn't enough willingness to run.	W
RV [nv]	*lehnt sich wieder zurück*				*Gestik im Folgenden weniger aggressiv und hektisch, d*	
WH [k]						

Figure 3: EXAKT

2.3. EXAKT

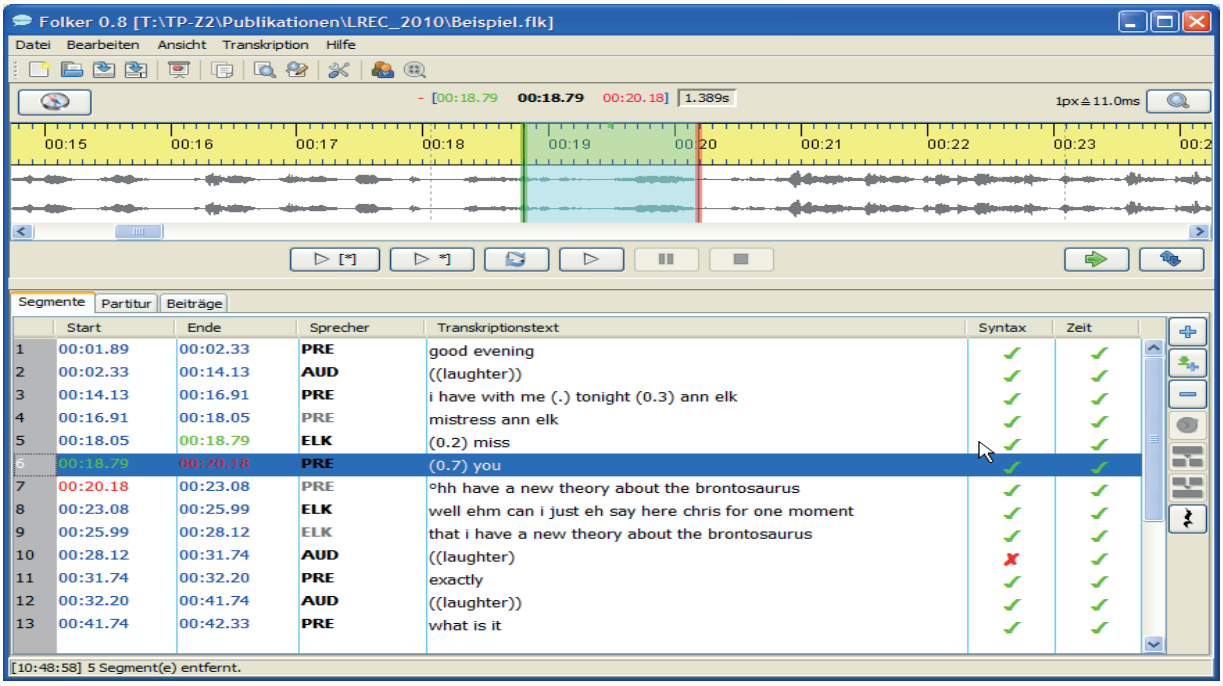
EXAKT ('EXMARaLDA Analysis- and Concordance tool') is a tool for querying one or several EXMARaLDA corpora which are represented in local or remote file systems or in a relational database. The basic functionality of the tool is that of a classical KWIC (Keyword in Context) concordancer. The user first formulates a query to the corpus. Queries can be carried out on the transcription data itself and/or on additional annotations referring to that transcription data. EXAKT offers several options for the query syntax, of which regular expressions are the most frequently used. The response to a query is a KWIC table containing all the matching annotations in the corpus. Starting from this table, the user can call up additional information of different kinds:

- Metadata assigned to the communication or to the speaker in question, for instance the age of the speaker at the time of recording. Such information will be displayed in additional columns of the KWIC table.
- The whole interactional context of the utterance in question, i.e. descriptions of what other speakers said or did during or around the same time. This information will be displayed as a transcription (in musical score or line notation) beneath the KWIC table when the user double clicks on a line of the concordance.
- The corresponding part of the underlying audio or video recording. This will be played back when the user selects a portion of the transcription underneath the KWIC table.
- Very frequently, the information contained in the corpus itself is not sufficient for an analysis, and users want to classify their search results according to additional categories. For that purpose, EXAKT provides so-called analysis columns which can either be filled with items from a closed list (e.g. a set of POS tags), with binary yes/no values (e.g. to distinguish analyzable from unclear cases) or with free text.

A KWIC table in EXAKT can thus be enriched with any number of additional columns containing different kinds of metadata or user-defined categorizations. Sorting the table according to one or several such columns is one way of discovering patterns or (ir)regularities in a query result. Another way is to further filter individual columns for certain properties, for instance: filter out those results that have been categorized as a proper name and/or those that are uttered by a speaker older than three years. Compared to a query in which all such parameters have to be specified in advance, such a stepwise process has the advantage of letting the user gradually encircle the phenomena he is interested in and thus minimizing the risk of systematically overlooking unanticipated empirical facts.

3. FOLK tools

FOLK (Forschungs- und Lehrkorpus, Deppermann and Hartung 2011) 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. The FOLKER transcription tool and the OrthoNormal annotation tool are one result of the work of that group.



	Start	Ende	Sprecher	Transkriptionstext	Syntax	Zeit
1	00:01.89	00:02.33	PRE	good evening	✓	✓
2	00:02.33	00:14.13	AUD	((laughter))	✓	✓
3	00:14.13	00:16.91	PRE	i have with me (.) tonight (0.3) ann elk	✓	✓
4	00:16.91	00:18.05	PRE	mistress ann elk	✓	✓
5	00:18.05	00:18.79	ELK	(0.2) miss	✓	✓
6	00:18.79	00:20.18	PRE	(0.7) you	✓	✓
7	00:20.18	00:23.08	PRE	°hh have a new theory about the brontosaurus	✓	✓
8	00:23.08	00:25.99	ELK	well ehm can i just eh say here chris for one moment	✓	✓
9	00:25.99	00:28.12	ELK	that i have a new theory about the brontosaurus	✓	✓
10	00:28.12	00:31.74	AUD	((laughter)	✗	✓
11	00:31.74	00:32.20	PRE	exactly	✓	✓
12	00:32.20	00:41.74	AUD	((laughter))	✓	✓
13	00:41.74	00:42.33	PRE	what is it	✓	✓

[10:48:58] 5 Segment(e) entfernt.

Figure 4: FOLKER Segment view

3.1. FOLKER

FOLKER – the FOLK Editor – is a transcription tool optimized for the workflow of the FOLK project. It uses the same codebase and basic data model as EXMARaLDA, but – in contrast to EXMARaLDA – privileges efficiency of transcription and annotation over diversity of application, i.e. it consciously reduces and rearranges the functionality available in EXMARaLDA in order to speed up and simplify the transcription process. 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.

The segment view (see 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 cGAT transcription conventions used for FOLK. Errors with respect to these conventions as well as violations of the temporal integrity of annotations (such as two overlapping annotations assigned to the same speaker) are indicated to the transcriber and can thus be fixed easily.

The Partitur (musical score) view displays the same transcription in a horizontal layout, organized into tiers. This view, which is basically identical to the graphical user interface of the EXMARaLDA Partitur-Editor (see figure 1), 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.

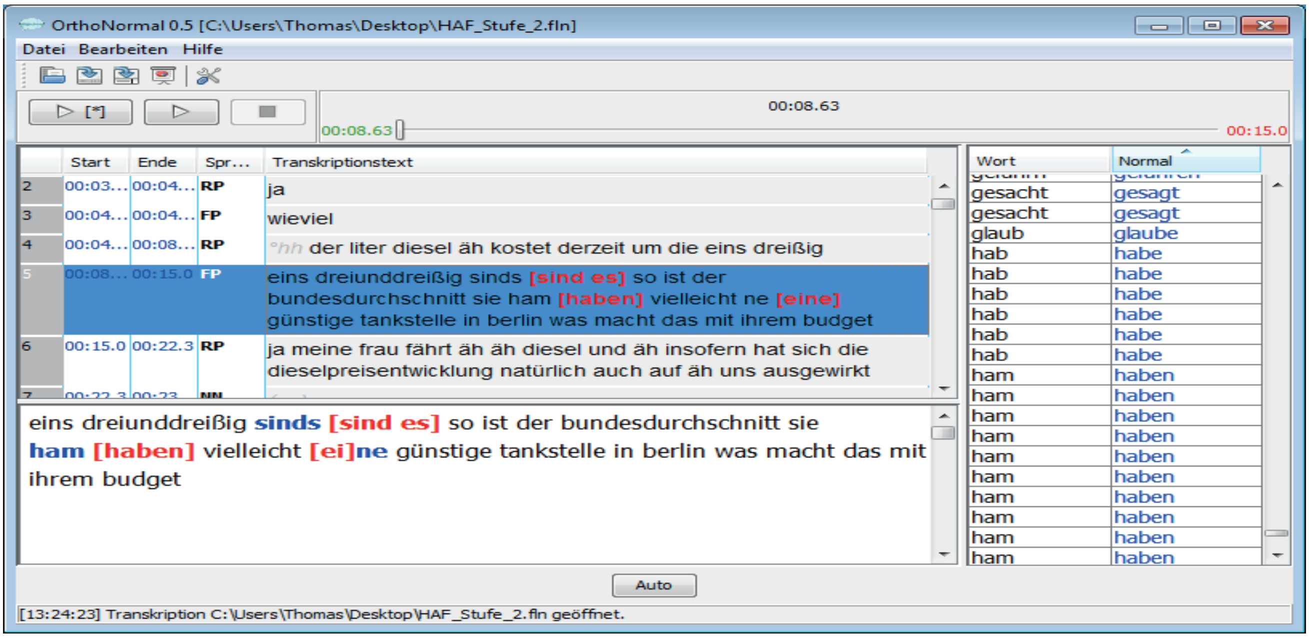
The contribution view, finally, also uses a vertically organized layout, but, instead of individual annotations, displays adjacent annotations of speakers as contributions. This view is thus close to a traditional, drama-script like representation of an interaction, complying with established reading 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.

3.2. OrthoNormal

Recordings in FOLK are transcribed according to the cGAT conventions which use so-called literary transcription to represent pronunciation deviations by means of orthographic symbols. Literary transcription has the advantage of being easy to apply by transcribers and easy to read by non-experts, but it also makes automatic processing and reliable querying of the data more difficult.

In order to overcome these difficulties, completed FOLK transcriptions are therefore orthographically normalized before they are integrated into the corpus database. Orthographic normalization means annotating all forms of literary transcription with a standard orthographic form. This process, although highly repetitive, has to be carried out manually because there are, as yet, no automatic algorithms for it.

The OrthoNormal tool (see figure 5) was developed to support the process of orthographic normalization as efficiently and consistently as possible. It takes a transcription file written by FOLKER as an input and allows the user to attach a normalized form to each word. Annotation speed and efficiency are increased through different means. First, the transcription is presented in two different views – a contribution view largely identical to that of the FOLKER tool, and a word list view in which all tokens can be sorted alphabetically. The latter view allows the user to annotate many identical forms in a single step. Second, annotation itself is carried out in a graphical component in which the input or selection of a normalized form can be done with only one or two mouse clicks. Third, a lookup in a lexicon in which all pairs of source and target forms are recorded with their frequencies is used to generate a list of suggestions for the annotator. As annotation of the corpus progresses and the lexicon grows, more and more annotations can thus be done as a selection from a list rather than through manual input of the normalized form.



	Start	Ende	Spr...	Transkriptionstext
2	00:03...	00:04...	RP	ja
3	00:04...	00:04...	FP	wieviel
4	00:04...	00:08...	RP	°hh der liter diesel äh kostet derzeit um die eins dreißig
5	00:08...	00:15.0	FP	eins dreiunddreißig sinds [sind es] so ist der bundesdurchschnitt sie ham [haben] vielleicht ne [eine] günstige tankstelle in berlin was macht das mit ihrem budget
6	00:15.0	00:22.3	RP	ja meine frau fährt äh äh diesel und äh insofern hat sich die dieselpreisentwicklung natürlich auch auf äh uns ausgewirkt

eins dreiunddreißig sinds [sind es] so ist der bundesdurchschnitt sie ham [haben] vielleicht [ei]ne günstige tankstelle in berlin was macht das mit ihrem budget

Wort	Normal
gesacht	gesagt
gesacht	gesagt
glaub	glaube
hab	habe
hab	habe
hab	habe
hab	habe
hab	habe
hab	habe
ham	haben
ham	haben
ham	haben
ham	haben
ham	haben
ham	haben
ham	haben
ham	haben
ham	haben
ham	haben

[13:24:23] Transkription C:\Users\Thomas\Desktop\HAF_Stufe_2.fln geöffnet.

Figure 5: OrthoNormal

4. Relation to other tools and tool sets

EXMARaLDA started out as a tool set for specific research community, attempting to cover its specific needs and requirements. Over the years, the system has grown in terms of functionality and has attracted users also from other fields of research than the one originally targeted. Thus, the system is used today also for research into dialectology, for annotation of written language, or for the construction of phonology corpora. Of course, new user communities have different and additional requirements which, in turn, lead to further extensions of the tools' functionality. The fact that most of these requirements could be satisfied proves the flexibility of the approach chosen. However, we also observe that a system which keeps growing in complexity also becomes harder to learn and use, especially for novices. The need for the development of FOLKER as a consciously reduced and simplified version of EXMARaLDA indicates that there are certain limits to what a single tool set can accomplish and that, at some point in the development, a choice has to be made between flexibility and efficiency.

If no single tool or tool set can satisfy all user requirements in an optimal way, interfaces between the tools must be developed which allow for an easy and, ideally, lossless exchange of data. EXMARaLDA and FOLKER, for instance, are perfectly interoperable so that corpus constructors can start with FOLKER for a quick and efficient basic transcription and then import these data into EXMARaLDA for more complex annotation tasks. EXMARaLDA, in turn, has import and export filters for almost all popular transcription tools, such as Praat, ELAN, CLAN, ANVIL, WinPitch or Transcriber, and also supports data exchange with the Audacity Audio Editor and TreeTagger, allowing for complex processing chains in which many different tools are used for different purposes. In the long run, however, interoperability should be guaranteed not through mechanism for exchange between individual tools, but rather through conformance with a generally accepted standard. Schmidt (2011) formulates a proposal for such a standard on the basis of the TEI guidelines. It will be one of the most important challenges for the speech and spoken language community to further the development of such proposals and to integrate standards into the workflows of the different research communities.

5. References

[1] Bird, S. and Liberman, M.. “A formal framework for linguistic annotation”, Speech Communication 33: 23-60, 2001.

[2] Deppermann, A. and Hartung, M., „Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) am Institut für Deutsche Sprache (Mannheim)“, to appear in: Felder, E., Müller, M.,Vogel, F. (eds.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen. Berlin, New York: de Gruyter.

[3] Rehbein, J., Schmidt, T., Meyer, B., Watzke, F. and Herkenrath, A., „Handbuch für das computergestützte Transkribieren nach HIAT“, Working Papers in Multilingualism, Series B 56, 2004.

[4] Schmidt, T., „A TEI-based approach to standardising spoken language transcription“, Journal of the Text Encoding Initiative (1), 2011.

[5] Schütte, W., „Transkriptionsrichtlinien für die Eingabe in EXMARaLDA (ab Version 1.2.7) nach DIDA-Konventionen.“ Mannheim: Institut für Deutsche Sprache, 2004.

[6] Selting, M., Auer, P. et al., “Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)”, Gesprächsforschung (10), 353-402, 2009.

[7] MacWhinney, B., “The CHILDES project: tools for analyzing talk.” Mahwah, NJ: Lawrence Erlbaum, 2000.