Erschienen in: Corpus linguistic software tools, Journal for language technology and computational linguistics Jg. 31 (2016) Nr. 1, S. 127-154

Thomas Schmidt

Construction and Dissemination of a Corpus of Spoken Interaction - Tools and Workflows in the FOLK project

Abstract

This paper is about the workflow for construction and dissemination of FOLK (Forschungs- und Lehrkorpus Gesprochenes Deutsch – Research and Teaching Corpus of Spoken German), a large corpus of authentic spoken interaction data, recorded on audio and video. Section 2 describes in detail the tools used in the individual steps of transcription, anonymization, orthographic normalization, lemmatization and POS tagging of the data, as well as some utilities used for corpus management. Section 3 deals with the DGD (Datenbank für Gesprochenes Deutsch - Database of Spoken German) as a tool for distributing completed data sets and making them available for qualitative and quantitative analysis. In section 4, some plans for further development are sketched.

1 Introduction

FOLK, the Forschungs- und Lehrkorpus Gesprochenes Deutsch (Research and Teaching Corpus of Spoken German) is being constructed in the program area "Oral Corpora" of the Institute for the German Language (IDS). Recognizing the lack of a larger, publicly available digital resource for studying spoken German in interaction (Deppermann/Hartung 2010), FOLK was started in 2009 as a long-term project to compile a diverse and systematic collection of audio and video recordings of spontaneous, authentic interactions across the whole spectrum of verbal interaction in German society.

FOLK is growing steadily, both in terms of quantity and variety of transcribed interactions. In its latest version (April 2016), the corpus comprises 219 interactions corresponding to 169 hours of audio and video recordings and 1.6 million transcribed word tokens. As testified by close to 6000 registrations (January 2017) for the Database of Spoken German (DGD, Schmidt 2014) through which FOLK is distributed and in which it is by far the most used corpus (Fandrych et al. 2016), the research community shows great interest in this resource.

To maximize (re)usability of the data, FOLK follows (and partly helps to define) current best practices in the handling and processing of data with respect to technological, methodological and legal issues (see also Schmidt 2016). In this paper, I am going to concentrate on the technological instruments, more specifically the software tools, which are used in the different steps of the corpus construction and dissemination workflow. Since FOLK is a spoken language corpus, the individual tools that make up this workflow differ fundamentally from the tools used in the compilation of written language corpora. Most importantly, the "primary" data of FOLK cannot be acquired automatically: recording authentic interactions requires an appropriate field access which must be negotiated for each individual case, and after data from the field have been obtained, project members have to screen and assess, and finally transcribe, them “manually”[1]. As described in Kupietz/Schmidt (2015), these two “bottlenecks” – field access and transcription – still prevent oral corpora from growing to the same dimensions as written corpora, and the transcription bottleneck makes up a great part of the technological challenges which FOLK faces.

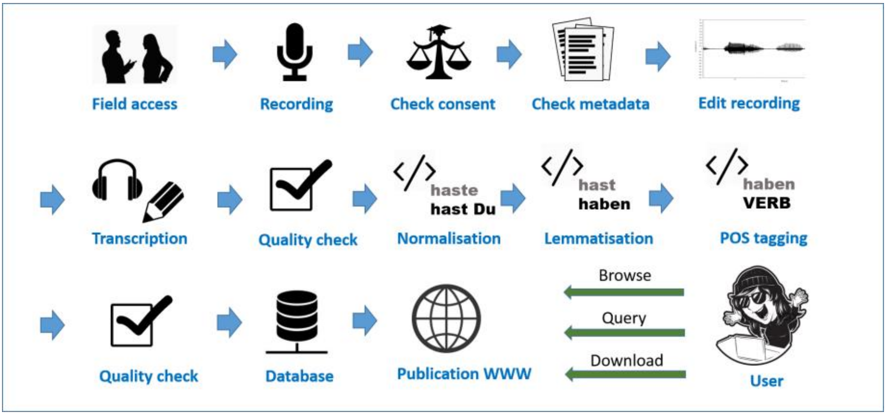


Figure 1: FOLK workflow overview

Figure 1 depicts (a simplified version of) the corpus compilation workflow in FOLK from the moment of field access to the final step of data publication. As a matter of principle, we do not start work on the data until the project coordinator has bindingly confirmed that all consent and metadata forms belonging to the recorded interaction are complete and usable. Once the data have passed this "gatekeeper" stage, recordings are prepared in the project's media studio for transcription. Depending on the recording, this step involves an appropriate conversion, cutting, denoising and/or normalization of the audio file as well as a synchronisation of different media files in cases where an interaction has been recorded in more than one file. Standard professional ("commercial") audio and video editing software (such as Samplitude and Adobe Premiere) is used for these tasks, which shall not be described in further detail here. Specialized linguistic tools come into play once an edited recording has been distributed to a student assistant for transcription.

Section 2 of this contribution starts at this stage, describing in detail the tools used in the individual steps of transcription, anonymization, orthographic normalization, lemmatization and POS tagging of the data, as well as some utilities used for corpus management. Section 3 then deals with the Database of Spoken German as a tool for distributing completed data

[1] It has been noted (p.c.) that “intellectually” may be the more appropriate term in this context, since it is the researcher’s informed involvement with the material (judging the authenticity and quality of a recording, taking interpretative decisions in the transcription process, etc.), rather than pure manual labour, that is decisive. I am using the word “manual” here because it is the most commonly used antonym to “automatic” when speaking about processing methods for language resources.

sets and making them available for qualitative and quantitative analysis. In section 4, some plans for further development are sketched.

2 Tools for Corpus Compilation

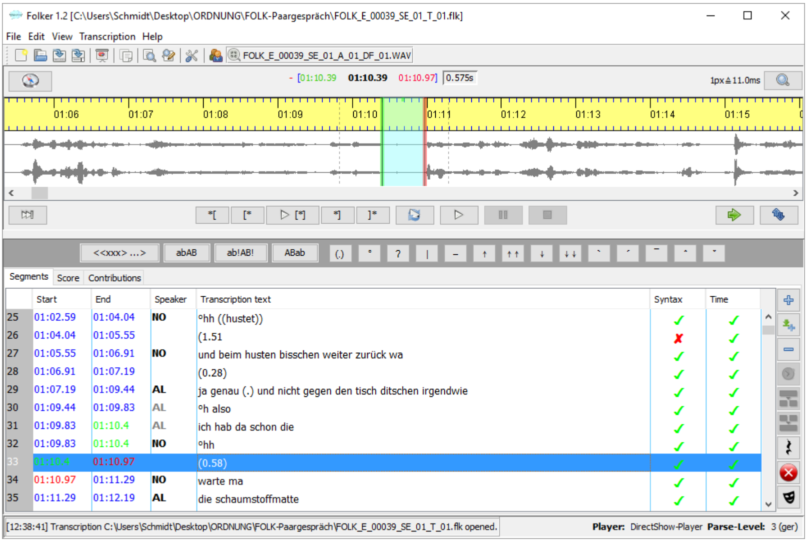
2.1 Transcription and Anonymization: FOLKER

FOLKER – the FOLK EditoR (see also Schmidt/Schütte 2010) – is based on, and to a great part reuses data models and code of, the EXMARaLDA system (Schmidt/Wörner 2014). The crucial difference to its "mother system" is that FOLKER is optimized for the particular task of transcription in the FOLK project. This means, first, that functionality not required in FOLK (such as manual alignment of existing legacy transcripts or free annotation in an arbitrary number of dependent tiers) is removed, thus reducing the complexity of the user interface, making it quicker and easier to learn for student transcribers and decreasing the number of opportunities for making errors in the transcription process. Second, in contrast to the EXMARaLDA Partitur-Editor (and similar tools like ELAN or Praat) which always presents data in a musical score view, FOLKER offers three different forms of data visualisation, each of which is particularly suitable for a specific step in the workflow. Third, FOLKER has direct built-in support for the transcription guideline of the FOLK project, the cGAT system. The following sections describe the functionality of FOLKER in more detail.

2.1.1 Transcription

Transcription in FOLK is done according to the guidelines for the cGAT minimal transcript (Schmidt et al. 2015). cGAT is based on the GAT2 system (Selting et al. 2009) which can be considered one of the most widely established transcription conventions in (German) conversation analysis and related fields. A cGAT minimal transcript requires careful transcription of individual words in modified orthography ("literarische Umschrift" – literary transcription, sometimes also referred to as "eye dialect"), a precise measurement of silences above 0.2s duration and a description of audible non-verbal interaction phenomena (like breathing, coughing or laughing). Aiming at a minimization of interpretative decisions in transcription, the cGAT minimal transcript does not require the identification of linguistic units above the word level (such as intonation phrases), the annotation of prosodic details (like primary accent or lengthening of syllables, voice quality, speed and modulation of speech) or comments interpreting individual parts of utterances (such as "ironic").

The initial transcription according to these guidelines is done in FOLKER's segment view (see figure 2). FOLK transcribers select suitable segments, typically of 2 to 5 seconds duration, in the wave form visualisation of the audio recording, and create a time-aligned segment of the recording (start and end times in the first and second column) which is assigned to a speaker (third column) and for which the actual transcription text can be entered (column 4). During transcription, this text is checked for formal compliance with the cGAT conventions. If an error is detected (such as the missing closing bracket for the pause in segment 26), this is indicated by a red cross (column 5), otherwise a green check mark confirms to the transcriber that the text entered can be parsed according to cGAT. Likewise, each segment is checked for “self-overlaps” with other segments, i.e. a red cross would indicate (in column 6) whenever the time intervals corresponding to two segments assigned to the same speaker overlap. All properties of a segment are freely editable at any time in the transcription process: time alignment can be adjusted, speaker assignment corrected, and transcription text modified whenever necessary.


Folker 1.2 [C:\Users\Schmidt\Desktop\ORDNUNG\FOLK-Paargespräch\FOLK_E_00039_SE_01_T_01.flk]
File Edit View Transcription Help
FOLK_E_00039_SE_01_A_01_DF_01.WAV
- [01:10.39 01:10.39 01:10.97] 0.575s
1px≙11.0ms
01:06 01:07 01:08 01:09 01:10 01:11 01:12 01:13 01:14 01:15
[[▷ [*] *]]*
<<xxx> ...> abAB ab!AB! ABab (.) ° ? | – ↑ ↑↑ ↓ ↓↓
Segments Score Contributions

	Start	End	Speaker	Transcription text	Syntax	Time
25	01:02.59	01:04.04	NO	°hh ((hustet))	✓	✓
26	01:04.04	01:05.55		(1.51	✗	✓
27	01:05.55	01:06.91	NO	und beim husten bisschen weiter zurück wa	✓	✓
28	01:06.91	01:07.19		(0.28)	✓	✓
29	01:07.19	01:09.44	AL	ja genau (.) und nicht gegen den tisch ditschen irgendwie	✓	✓
30	01:09.44	01:09.83	AL	°h also	✓	✓
31	01:09.83	01:10.4	AL	ich hab da schon die	✓	✓
32	01:09.83	01:10.4	NO	°hh	✓	✓
33	01:10.4	01:10.97		(0.58)	✓	✓
34	01:10.97	01:11.29	NO	warte ma	✓	✓
35	01:11.29	01:12.19	AL	die schaumstoffmatte	✓	✓

[12:38:41] Transcription C:\Users\Schmidt\Desktop\ORDNUNG\FOLK-Paargespräch\FOLK_E_00039_SE_01_T_01.flk opened.

Player: DirectShow-Player **Parse-Level:** 3 (ger)

Figure 2: FOLKER's segment view

Although FOLKER is not meant for video transcription proper – meaning the systematic annotation of (non-verbal) behaviour visible in video images – a video file can be loaded into the editor in addition to the audio file to facilitate speaker identification and the understanding of passages with no or little verbal output (figure 3).

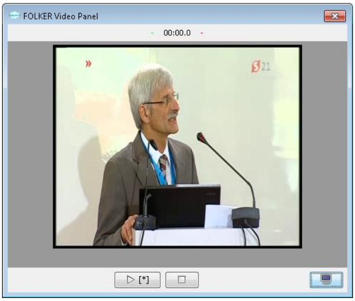
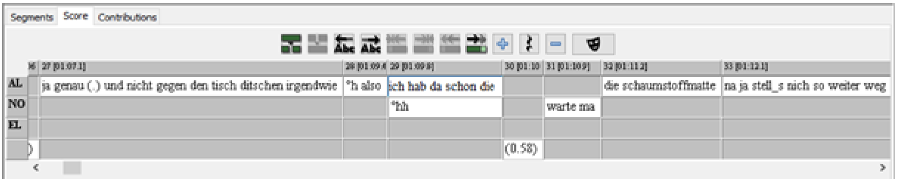


Figure 3: FOLKER's video panel

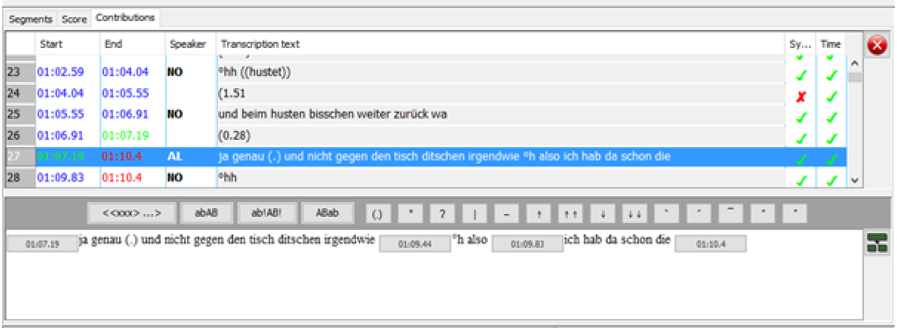
In cases of simultaneous or overlapping speech (a ubiquitous property of the types of interaction included in FOLK), transcribers first create independent segments for each speaker. The precise specification of the start and end of an overlap of one speaker's segment in relation to another speaker's segment can then be carried out by switching to FOLKER's musical score ("Partitur") view whose two-dimensional layout presents temporal relations in a more intuitive way than the segment view (see figure 4).



	36	27 [01:07.1]	28 [01:09.4	29 [01:09.8]	30 [01:10	31 [01:10.9]	32 [01:11.2]	33 [01:12.1]
AL		ja genau (.) und nicht gegen den tisch ditschen irgendwie	°h also	ich hab da schon die			die schaumstoffmatte	na ja stell_s nich so weiter weg
NO				°hh		warte ma		
EL								
	)				(0.58)			

Figure 4: FOLKER's musical score ("Partitur") view

While transcription itself is thus done in the segment and musical score view with transcribers freely switching between the two as appropriate, a third view is used for intermediate or concluding quality checks on the data. In the contribution view (figure 5), consecutive segments assigned to the same speaker are summarized into the larger unit of a speaker contribution. This visualisation makes it easier to read the transcription as a whole and thus makes the correction process more efficient.



	Start	End	Speaker	Transcription text	Sy...	Time
23	01:02.59	01:04.04	NO	°hh ((hustet))	✓	✓
24	01:04.04	01:05.55		(1.51	✗	✓
25	01:05.55	01:06.91	NO	und beim husten bisschen weiter zurück wa	✓	✓
26	01:06.91	01:07.19		(0.28)	✓	✓
27	01:07.19	01:10.4	AL	ja genau (.) und nicht gegen den tisch ditschen irgendwie °h also ich hab da schon die	✓	✓
28	01:09.83	01:10.4	NO	°hh	✓	✓

01:07.19 ja genau (.) und nicht gegen den tisch ditschen irgendwie 01:09.44 °h also 01:09.83 ich hab da schon die 01:10.4

Figure 5: FOLKER's contribution view

2.1.2 Data format

FOLKER reads and writes an XML data format in which all relevant entities of the transcription – recordings, speakers, timepoints, speaker contributions – are represented as elements, and their relationships – speaker assignment, temporal alignment – encoded via IDREF/ID pointers. For a transcript which follows in its entirety the conventions for the cGAT minimal transcript (i.e. for which a transcriber sees nothing but green check marks in the respective columns of the editor), FOLKER can parse the transcription text, resulting in additional markup of word, pause, breathing etc. elements underneath the speaker contributions. Figure 6 shows the XML corresponding to the transcript excerpt used in the previous section.

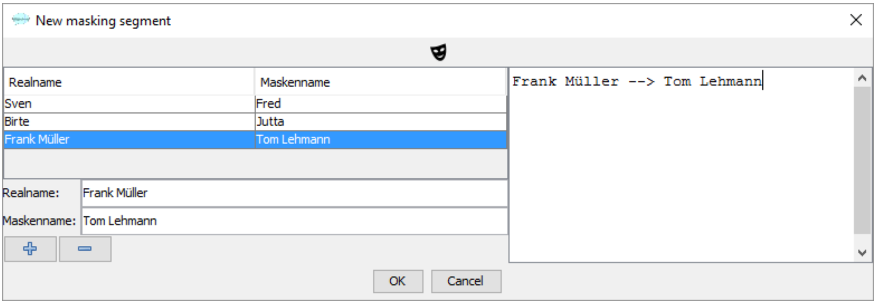
```xml
<speakers>
    <speaker speaker-id="NO"><name>Norbert</name></speaker>
    <speaker speaker-id="EL"><name>Elena</name></speaker>
</speakers>

<recording path="FOLK_E_00039_SE_01_A_01_DF_01.WAV"/>

<timeline>
    <timepoint timepoint-id="TLI_0" absolute-time="0.0"/>
    <!-- [...] -->
    <timepoint timepoint-id="TLI_25" absolute-time="65.555"/>
    <timepoint timepoint-id="TLI_26" absolute-time="66.915"/>
    <timepoint timepoint-id="TLI_27" absolute-time="67.195"/>
    <timepoint timepoint-id="TLI_28" absolute-time="69.44"/>
    <timepoint timepoint-id="TLI_29" absolute-time="69.83"/>
    <timepoint timepoint-id="TLI_30" absolute-time="70.4"/>
    <!-- [...] -->
</timeline>

<contribution speaker-reference="NO" start-reference="TLI_25" end-reference="TLI_26">
    <w>und</w><w>beim</w><w>husten</w><w>bisschen</w>
    <w>weiter</w><w>zurück</w><w>wa</w>
</contribution>
<contribution start-reference="TLI_26" end-reference="TLI_27" parse-level="2">
    <pause duration="0.28"/>
</contribution>
<contribution speaker-reference="AL" start-reference="TLI_27" end-reference="TLI_30">
    <w>ja</w><w>genau</w>
    <pause duration="micro"/>
    <w>und</w><w>nicht</w><w>gegen</w><w>den</w>
    <w>tisch</w><w>ditschen</w><w>irgendwie</w>
    <time timepoint-reference="TLI_28"/>
    <breathe type="in" length="1"/>
    <w>also</w>
    <time timepoint-reference="TLI_29"/>
    <w>ich</w><w>hab</w><w>da</w><w>schon</w><w>die</w>
</contribution>
```

Figure 6: FOLKER's XML format

The FOLKER XML format, in its unparsed as well as in its parsed version, and also with additional lemma and POS information for the tokens (see sections 2.2 and 2.3), is isomorphic and can be easily transformed to the TEI-based ISO standard "Transcriptions of Spoken Language" (ISO/ TC 37/SC 4/WG 6, cf. Schmidt/Hedeland/Jettka 2017), published in August 2016. FOLKER as well as OrthoNormal offer export filters, and the DGD enables users to download FOLK excerpts in this format (see section 3.3). In future developments, we will make sure to maintain interoperability with the ISO standard, and, eventually, rebase the whole FOLK workflow on it.

2.1.3 Anonymization / Pseudonymization / Masking

Recordings in FOLK are done with informed consent of the speakers wherever this is legally required (i.e. almost always). The standard consent form (for audio recordings[2]) guarantees that all information that could lead to direct identification of an individual, in particular the mention of individuals' names, addresses or other specific biographic details, are replaced in metadata and transcriptions with suitable pseudonyms and in the recordings with a silence or noise. Identifying the places to be masked in the recordings and maintaining a consistent set of pseudonyms for use in the transcription is another laborious task requiring adequate tool support. In most cases, the best moment to decide on anonymization issues is during the transcription process itself, when transcribers carefully listen to the recordings anyway and can thus be sure to notice mentions of proper names etc. FOLKER therefore includes a set of functions which support transcribers in this task. Whenever an anonymization issue is identified, the corresponding part of the recording can be selected and a masking segment created which is stored separately from the transcription. In order to ensure consistency in the choice of pseudonyms (i.e. to make sure that one and the same name is always replaced by one and the same pseudonym), transcribers can create and maintain a table of mappings from real names to pseudonyms (see figure 7).



Realname	Maskenname
Sven	Fred
Birte	Jutta
Frank Müller	Tom Lehmann

Realname: Frank Müller

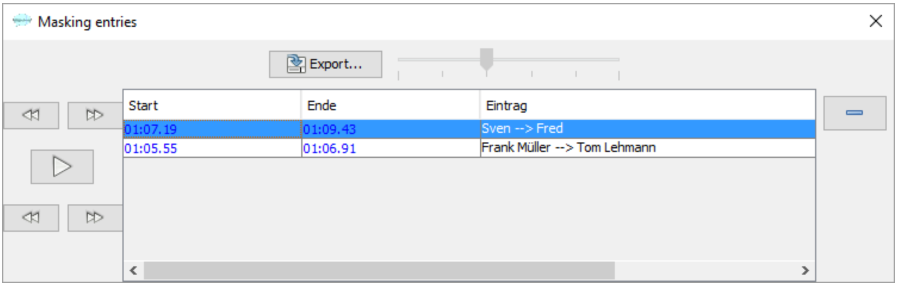
Maskenname: Tom Lehmann

```
Frank Müller --> Tom Lehmann
```

Figure 7: Creating a mask segment and managing pseudonyms in FOLKER

Since anonymization is, ultimately, not a decision for which student transcribers can or should take full responsibility, completed transcripts and the anonymizations proposed by the transcribers are checked by a project coordinator. For this task, FOLKER offers a summary of all existing masking segments as illustrated in figure 8.

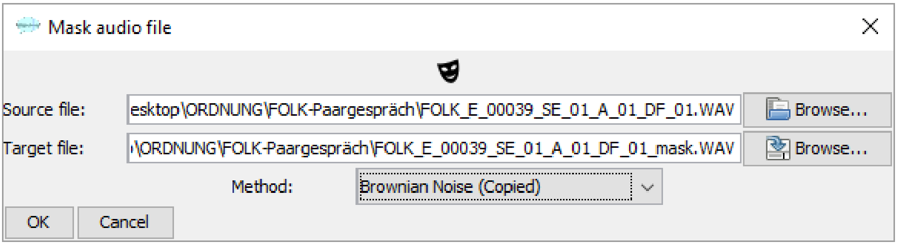
[2] The same anonymization principles apply to the sound track of video recordings. We do not, however, attempt to anonymize the video image, for instance by blurring faces, since this would render the video useless for many analysis purposes. Instead, we obtain the speakers' consent to use the unmasked video image.



Start	Ende	Eintrag
01:07.19	01:09.43	Sven --> Fred
01:05.55	01:06.91	Frank Müller --> Tom Lehmann

Figure 8: Overview of masking segments in FOLKER

When all masking segments for a given transcript have been identified and checked in this way, FOLKER can insert the required noises into the audio file automatically (figure 9). We use a Brownian noise, because, in contrast to a simple silence, this makes clear to the listener that he is dealing with an artefact in the recording, and because, compared to a white noise, it is less disagreeable to the ear. Masking information is stored in a separate section of the document so that it can be easily removed before publication of the data. We archive this information internally in order to be able to quickly identify masked passages later.



Source file: esktop\ORDNUNG\FOLK-Paargespräch\FOLK_E_00039_SE_01_A_01_DF_01.WAV

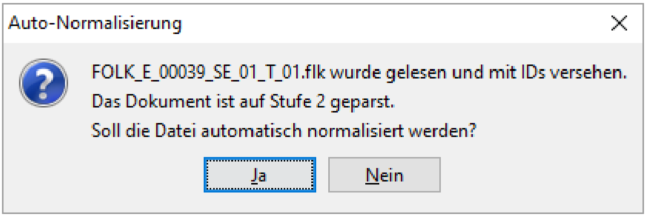
Target file: \ORDNUNG\FOLK-Paargespräch\FOLK_E_00039_SE_01_A_01_DF_01_mask.WAV

Method: Brownian Noise (Copied)

Figure 9: Automatic masking of an audio file via FOLKER

2.2 Orthographic normalisation: OrthoNormal

As described above, primary transcription in FOLK is done according to cGAT, meaning that all word tokens are written in lower case, and that deviations from standard pronunciation are modelled by using a modified orthography (e.g. “zwo” as a colloquial pronunciation of the number 2 or “haste” as a contracted form of “hast Du”, have you – “dunno” for “don’t know” would be an analogous case for English). While this has the advantage of following conversation analytic tradition and making spoken language phenomena more readily visible in the transcription text, it also has the disadvantage of rendering queries and further automatic processing on this data unreliable. In order to optimize FOLK data for the application of corpus linguistic and computational linguistic methods, we therefore add a second annotation layer in which tokens in modified orthography are mapped to a standard orthographic equivalent.[3] This is done on a token-by-token basis with the help of the OrthoNormal annotation tool using a set of normalisation guidelines (Schmidt/Winterscheid 2015).


Auto-Normalisierung

FOLK_E_00039_SE_01_T_01.flk wurde gelesen und mit IDs versehen.
Das Dokument ist auf Stufe 2 geparst.
Soll die Datei automatisch normalisiert werden?

Figure 10: Automatic normalisation in OrthoNormal

When a completed FOLKER transcript is loaded into OrthoNormal, an automatic normalisation step can be applied to all word tokens (see figure 10). This method proceeds as follows:

1 It looks up each word in a normalisation lexicon in which (manually verified) transcription/normalisation pairs of previous normalisations are stored with their frequencies.[4] Whenever a form is encountered that has an entry in this lexicon, the most frequent corresponding normalised form is automatically inserted. As an example, see the form “hab” in figure 11 which has been correctly normalised to “habe” (*have*, first person singular), but also the form “wa” which has been incorrectly normalised to “wir” (*we*, where “was” – *what* would have been correct).

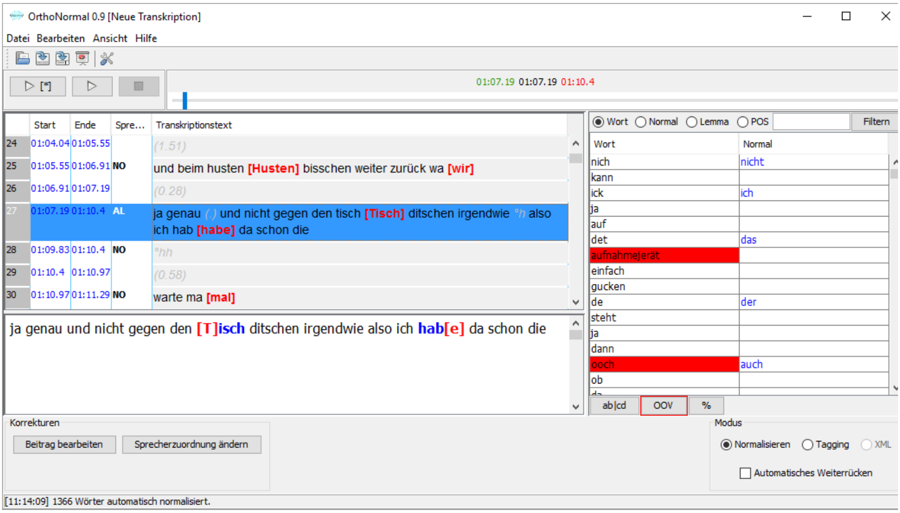
2 It looks up each word in a list of word forms that only occur in upper case in German, extracted from the DeReWo list of inflected forms (Institut für Deutsch Sprache 2014) which itself is based on the billion words DeReKo corpus of written German (Kupietz/Schmidt 2015). If no lexicon entry for a given form has been found in step (1) and the form with an upper-case initial is found in the word list, this upper case form is inserted as the normalised form. As an example, see the form “tisch” in figure 11 which has been correctly normalised to “Tisch” – *table*.

[3] As a reviewer has duly pointed out, other projects such as Verbmobil have proceeded in the reverse manner, i.e. standard orthography was used in the primary transcription and pronunciation deviations added as annotations to the orthographic words. Our choice to transcribe in modified orthography is mainly motivated by the fact that this is the standard procedure in conversation analysis and therefore more easily reconcilable with existing transcription conventions. Furthermore, FOLK data abound with dialectal and other features of spontaneous speech so that the rate of forms deviating from standard orthography is rather high (more than 50% of all tokens in some cases). A partial automation of the mapping between modified and standard forms is therefore important for reasons of efficiency, and it is obviously much easier to automatically map modified onto standard forms than the other way around.

[4] This lexicon is updated with each release of FOLK, i.e. it grows by the manually verified normalization entries for roughly 300.000 transcribed tokens each year.

3. It checks all word forms against a full list of inflected German word forms (again, based on DeReWo). If a form is not found in the list, it will be marked as a likely normalisation candidate for the manual normalisation process. As an example, see the forms “aufnahmejerät” (=”Aufnahmegerät” – *recording device*) and “ooch” (=”auch” – *too* – in Berlin dialect), both highlighted in red in the table on the right hand side of figure 11.

This simple process leads to recall and precision rates both roughly around 80%, meaning that 80% of forms that need to be normalised are detected in that process and that 80% of all automatically inserted normalised forms are correct. Since the normalisation layer is absolutely crucial for all further processing steps, the automatically normalised transcripts with this error rate are manually checked and corrected by student annotators. The OrthoNormal tool makes this step efficient by offering an ergonomic interface optimised for the task. As figure 11 illustrates, the interface is divided into three parts: the upper left part displays the transcript as a list of contributions with normalised forms added in red. When a contribution is selected in this list, the lower left part displays this contribution and makes it available for editing. Clicking on any word token in this view will bring up a normalisation dialog in which a normalised form can either be freely entered in a text field, or selected from a list of candidates extracted from the normalisation lexicon. The right part of the screen, finally, displays pairs of transcribed and normalised forms in a table. When ordered alphabetically, several transcribed forms can be normalised in one go in this table, and a regular expression filter can be used to select specific patterns of transcribed forms (such as: all forms starting with a certain prefix).


OrthoNormal 0.9 [Neue Transkription]
Datei Bearbeiten Ansicht Hilfe
▷ [*]
▷
01:07.19 01:07.19 01:10.4
● Wort ○ Normal ○ Lemma ○ POS
Filtern
ab|cd
OOV
%
Korrekturen
Beitrag bearbeiten
Sprecherzuordnung ändern
Modus
● Normalisieren ○ Tagging ○ XML
□ Automatisches Weiterrücken

	Start	Ende	Spre…	Transkriptionstext
24	01:04.04	01:05.55		(1.51)
25	01:05.55	01:06.91	NO	und beim husten [Husten] bisschen weiter zurück wa [wir]
26	01:06.91	01:07.19		(0.28)
27	01:07.19	01:10.4	AL	ja genau (.) und nicht gegen den tisch [Tisch] ditschen irgendwie °h also ich hab [habe] da schon die
28	01:09.83	01:10.4	NO	°hh
29	01:10.4	01:10.97		(0.58)
30	01:10.97	01:11.29	NO	warte ma [mal]

ja genau und nicht gegen den [T]isch ditschen irgendwie also ich hab[e] da schon die

Wort	Normal
nich	nicht
kann	
ick	ich
ja	
auf	
det	das
aufnahmejerät	
einfach	
gucken	
de	der
steht	
ja	
dann	
ooch	auch
ob	

[11:14:09] 1366 Wörter automatisch normalisiert.

Figure 11: OrthoNormal user interface

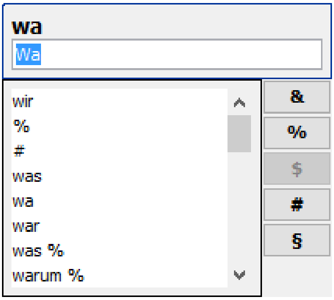
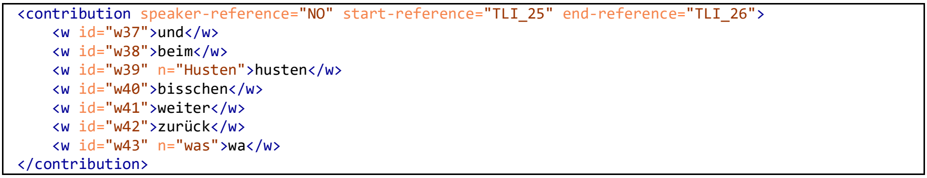


Figure 12: Normalisation dialog

Although manual normalisation means that a student annotator has to go through the whole transcript once more (and the result checked again by a supervisor), this step is by far less time-consuming than transcription itself. While we have to calculate with as much as 100 hours of manual work for the transcription of 1 hour of recording, the same amount of data can usually be orthographically normalised in less than 5 hours. The normalised forms are stored as @n attributes on the <w> elements of the original transcription (see figure 13).



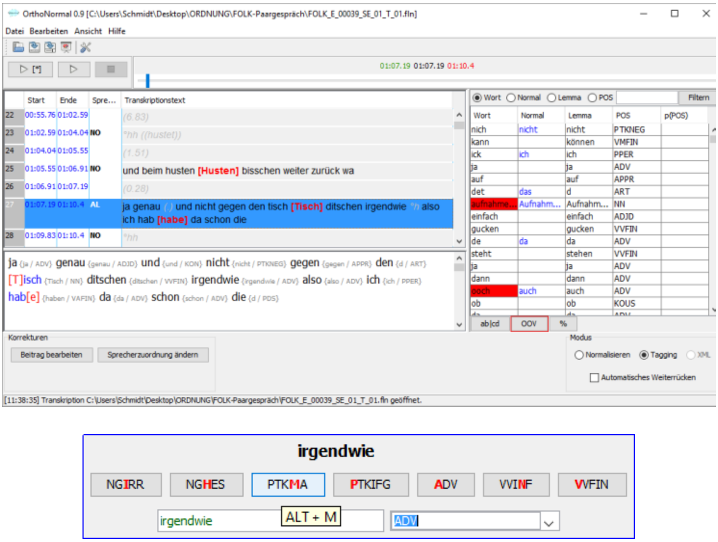
```
<contribution speaker-reference="NO" start-reference="TLI_25" end-reference="TLI_26">
    <w id="w37">und</w>
    <w id="w38">beim</w>
    <w id="w39" n="Husten">husten</w>
    <w id="w40">bisschen</w>
    <w id="w41">weiter</w>
    <w id="w42">zurück</w>
    <w id="w43" n="was">wa</w>
</contribution>
```

Figure 13: XML of normalised transcript

2.3 Lemmatisation and POS-Tagging

Once the manually checked normalisation layer is available for a transcript, an automatic lemmatization and POS tagging can be carried out on the normalised data. We use TreeTagger (Schmid 1994) via TT4J (Eckart de Castilho, no data) to perform this task.

Up until the current version of FOLK, the default TreeTagger parameter file for German, trained on newspaper text with the Stuttgart-Tübingen-Tagset (STTS), was used to do the tagging. The results were acceptable only as a first approximation, because they had an error rate of over 10% for POS tags (less than 2% for lemmas), and because the tagset itself was underspecified especially for those word classes that are specific to spoken language (such as particles and interjections). Westpfahl (2015) therefore developed an extension of STTS optimized for the kind of spoken language data found in FOLK. In the FOLK project, a 100.000 tokens gold standard (Westpfahl/Schmidt 2016) was tagged manually according to this STTS extension, again by means of the OrthoNormal tool (in the "tagging" rather than the "normalisation" mode, see figure 14). Using this gold standard, a new TreeTagger parameter file was trained which can be used for lemmatization and POS tagging of future versions of FOLK. Evaluations have shown that an error rate as low as 5% can be attained with this improved procedure.



OrthoNormal 0.9 [C:\Users\Schmidt\Desktop\ORDNUNG\FOLK-Paargespräch\FOLK_E_00039_SE_01_T_01.fln]

01:07.19 01:07.19 01:10.4

	Start	Ende	Spre...	Transkriptionstext
22	00:55.76	01:02.59		(6.83)
23	01:02.59	01:04.04	NO	°hh ((hustet))
24	01:04.04	01:05.55		(1.51)
25	01:05.55	01:06.91	NO	und beim husten [Husten] bisschen weiter zurück wa
26	01:06.91	01:07.19		(0.28)
27	01:07.19	01:10.4	AL	ja genau (.) und nicht gegen den tisch [Tisch] ditschen irgendwie °h also ich hab [habe] da schon die
28	01:09.83	01:10.4	NO	°hh

ja {ja / ADV} genau {genau / ADJD} und {und / KON} nicht {nicht / PTKNEG} gegen {gegen / APPR} den {d / ART} [T]isch {Tisch / NN} ditschen {ditschen / VVFIN} irgendwie {irgendwie / ADV} also {also / ADV} ich {ich / PPER} hab[e] {haben / VAFIN} da {da / ADV} schon {schon / ADV} die {d / PDS}

● Wort ○ Normal ○ Lemma ○ POS

Wort	Normal	Lemma	POS	p(POS)
nich	nicht	nicht	PTKNEG	
kann		können	VMFIN	
ick	ich	ich	PPER	
ja		ja	ADV	
auf		auf	APPR	
det	das	d	ART	
aufnahme...	Aufnahm...	Aufnahm...	NN	
einfach		einfach	ADJD	
gucken		gucken	VVFIN	
de	da	da	ADV	
steht		stehen	VVFIN	
ja		ja	ADV	
dann		dann	ADV	
ooch	auch	auch	ADV	
ob		ob	KOUS	

Modus: ○ Normalisieren ● Tagging ○ XML

☐ Automatisches Weiterrücken

[11:38:35] Transkription C:\Users\Schmidt\Desktop\ORDNUNG\FOLK-Paargespräch\FOLK_E_00039_SE_01_T_01.fln geöffnet.

irgendwie

irgendwie

ADV

Figure 14: Using OrthoNormal do carry out manual correction on POS tags

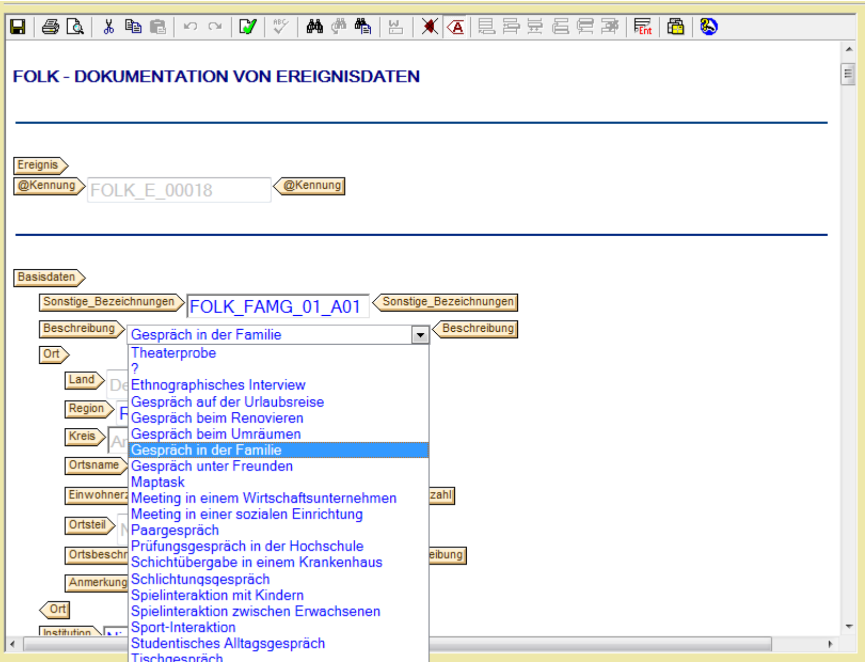
Lemmas and POS tags are, again, stored as @lemma and @pos attributes, respectively, on the <w> elements of the original transcription (see figure 15).

```
 <contribution speaker-reference="NO" start-reference="TLI_25" end-reference="TLI_26">
    <w id="w39" pos="KON" lemma="und">und</w>
    <w id="w40" pos="APPRART" lemma="beim">beim</w>
    <w id="w41" n="Husten" pos="NN" lemma="Husten">husten</w>
    <w id="w42" pos="ADV" lemma="bißchen">bisschen</w>
    <w id="w43" pos="ADV" lemma="weiter">weiter</w>
    <w id="w44" pos="PTKVZ" lemma="zurück">zurück</w>
    <w id="w45" pos="SEQU" lemma="Wa" n="was">wa</w>
</contribution>
```

Figure 15: XML of lemmatized and POS tagged transcript

2.4 Metadata

Metadata capturing salient characteristics of the interactions and speakers involved are collected alongside the recordings in the field by means of a project specific paper form. Once a recording is approved for inclusion in FOLK and the project coordinator has checked that consent and metadata forms for this recording are complete and consistent, metadata are transferred to a digital form. This is done with the help of an online interface based on the XMLSpy Editor (see figure 16). The interface is aware of the underlying XML schema (Gasch et al. 2008) and can thus support the entry process, for instance by providing closed vocabulary lists for values of appropriate fields.



FOLK - DOKUMENTATION VON EREIGNISDATEN

Ereignis

@Kennung FOLK_E_00018 @Kennung

Basisdaten

Sonstige_Bezeichnungen FOLK_FAMG_01_A01 Sonstige_Bezeichnungen

Beschreibung Gespräch in der Familie Beschreibung

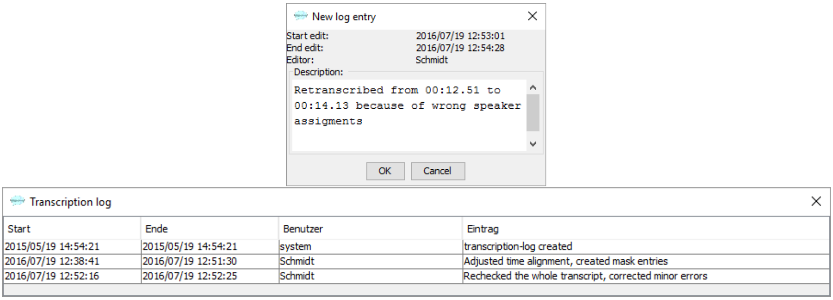
Ort
Land
Region
Kreis
Ortsname
Einwohnerz
Ortsteil
Ortsbeschr
Anmerkung
Ort
Institution
zahl
eibung
Theaterprobe
?
Ethnographisches Interview
Gespräch auf der Urlaubsreise
Gespräch beim Renovieren
Gespräch beim Umräumen
Gespräch in der Familie
Gespräch unter Freunden
Maptask
Meeting in einem Wirtschaftsunternehmen
Meeting in einer sozialen Einrichtung
Paargespräch
Prüfungsgespräch in der Hochschule
Schichtübergabe in einem Krankenhaus
Schlichtungsgespräch
Spielinteraktion mit Kindern
Spielinteraktion zwischen Erwachsenen
Sport-Interaktion
Studentisches Alltagsgespräch
Tischgespräch

Figure 16: Web interface for entering metadata

2.5 Data Management

Because of the large amount of manual work necessary for transcribing and annotating the data, the FOLK project continuously employs a team of 10 to 15 student assistants. This, and the fact that the acquisition of new recordings cannot be planned centrally, but has to be managed individually for each new type of interaction with the respective cooperation partners, lead to a considerable administrational overhead. As the project progresses, we are attempting to develop tools not only for transcription and annotation itself, but also for supporting the project management in monitoring the workflows.

1 In order to monitor progress on individual transcription files, FOLKER offers transcribers the possibility to keep a transcription log, a simple list of logging entries with information about the time in which a transcript was edited, the name of the person who did the editing, and a free text field describing the editing steps carried out (see figure 17). When aggregated over a larger number of files, this information can also be used to measure transcription ratios.



New log entry

Start edit: 2016/07/19 12:53:01
End edit: 2016/07/19 12:54:28
Editor: Schmidt
Description:

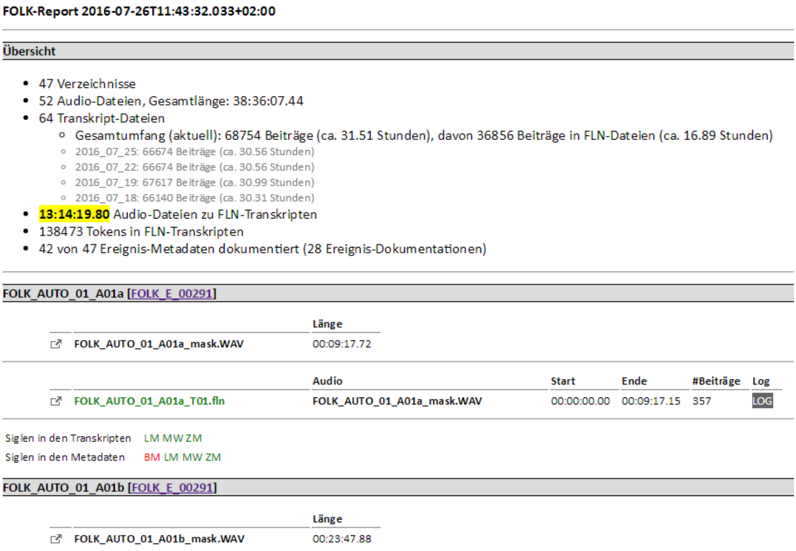
```
Retranscribed from 00:12.51 to
00:14.13 because of wrong speaker
assigments
```

Transcription log

Start	Ende	Benutzer	Eintrag
2015/05/19 14:54:21	2015/05/19 14:54:21	system	transcription-log created
2016/07/19 12:38:41	2016/07/19 12:51:30	Schmidt	Adjusted time alignment, created mask entries
2016/07/19 12:52:16	2016/07/19 12:52:25	Schmidt	Rechecked the whole transcript, corrected minor errors

Figure 17: Transcription logs in FOLKER

2 In order to monitor progress of transcription and normalisation on the corpus as a whole, a set of batch scripts has been implemented to produce so-called snapshots of the current state of corpus development. A single click will start a process which runs through all folders in the project's working directory, calculates measurements for transcription progress (e.g. amount of audio available, amount of audio transcribed, number of files normalized, number of metadata entries completed) and generates HTML visualizations for transcript logs and transcription files. For instance, the snapshot depicted in figure 18 informs the project coordinator that, out of a total of 38.5 hours of recordings, roughly 31.5 hours have been transcribed at least in a first pass, and a little more than 13 hours have already entered the normalisation stage.



FOLK-Report 2016-07-26T11:43:32.033+02:00

Übersicht

- 47 Verzeichnisse
- 52 Audio-Dateien, Gesamtlänge: 38:36:07.44
- 64 Transkript-Dateien
  - Gesamtumfang (aktuell): 68754 Beiträge (ca. 31.51 Stunden), davon 36856 Beiträge in FLN-Dateien (ca. 16.89 Stunden)
  - 2016_07_25: 66674 Beiträge (ca. 30.56 Stunden)
  - 2016_07_22: 66674 Beiträge (ca. 30.56 Stunden)
  - 2016_07_19: 67617 Beiträge (ca. 30.99 Stunden)
  - 2016_07_18: 66140 Beiträge (ca. 30.31 Stunden)
- **13:14:19.80** Audio-Dateien zu FLN-Transkripten
- 138473 Tokens in FLN-Transkripten
- 42 von 47 Ereignis-Metadaten dokumentiert (28 Ereignis-Dokumentationen)

FOLK_AUTO_01_A01a [FOLK_E_00291]

	Länge
FOLK_AUTO_01_A01a_mask.WAV	00:09:17.72

	Audio	Start	Ende	#Beiträge	Log
FOLK_AUTO_01_A01a_T01.fln	FOLK_AUTO_01_A01a_mask.WAV	00:00:00.00	00:09:17.15	357	LOG

Siglen in den Transkripten	LM MW ZM
Siglen in den Metadaten	BM LM MW ZM

FOLK_AUTO_01_A01b [FOLK_E_00291]

	Länge
FOLK_AUTO_01_A01b_mask.WAV	00:23:47.88

Figure 18: Report on progress in the transcription and annotation process

3 The workflow schema depicted in figure 1 is oversimplified in one important detail: it fails to capture cycles in the workflow that arise from the fact that transcriptions and

annotations of oral language data will always contain a portion of genuine errors. Some of these errors are discovered (by project members or users of the corpus) only after a piece of data has been declared complete and disseminated via the database. Although, owing to the different quality control steps in the workflow, this is rare, it is not rare enough to be simply ignored. An additional important part of the workflow is therefore the management of correction cycles. In order to control these cycles, we manage the transcription data (as well as the metadata, for which similar problems can occur) via Subversion (SVN) as a version control system and coordinate the yearly extension of FOLK with corrections that have in the meantime been applied to the already published part.

3. Tools for Corpus Analysis: Database for Spoken German (DGD)

The observation that "[...] a corpus by itself can do nothing at all, being nothing more than a store of used language" is no less true for oral corpora than for the written corpora Hunston (2002: 20) refers to. Corpus linguists need adequate tools not only for constructing but also for analysing corpora. In the case of FOLK, the Database for Spoken German (Datenbank für Gesprochenes Deutsch, DGD, http://dgd.ids-mannheim.de) is the principal means of making the corpus data available for analysis. The DGD in its present form (versions 2.x – following up on the predecessor system described in Fiehler/Wagner 2005), first released in 2012, acts as a platform not only for disseminating FOLK, but also various other oral corpora stored at the Archive for Spoken German (AGD).

The DGD allows for two principal approaches to oral corpus data:

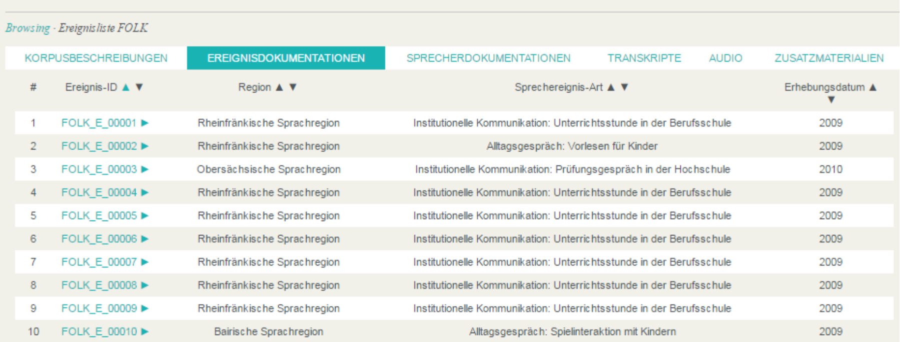
1 **Browsing** a corpus, i.e. reading corpus metadata and transcripts and listening to the corresponding (aligned) audio is a means of getting acquainted with a corpus, of exploring individual data sets in a holistic manner and of identifying and analysing in depth selected excerpts of transcripts. Related to this is the possibility of **downloading** selected excerpts and further processing them with suitable tools (e.g. for acoustic analysis, for additional annotations) on a local machine. The browsing approach is particularly suited for qualitative paradigms, such as conversation analysis.

2 **Querying** a corpus, i.e. searching through the entire data for all instances of a given annotation pattern and further manipulating and analysing the results of such searches. This is the functionality typically expected from a corpus interface to written language data, and it is equally central to the work with oral data. Corpus queries are essential for quantitative research paradigms, certain corpus linguistic methods being the most obvious case in point.

Of course, the real potential of a corpus like FOLK lies in an innovative mixture of these two approaches, and, as will be shown in the following sections, the DGD takes great care to enable users to effectively combine "semi-automatic" query methods with "manual" ways of inspecting the data.

3.1 Browsing / Collections

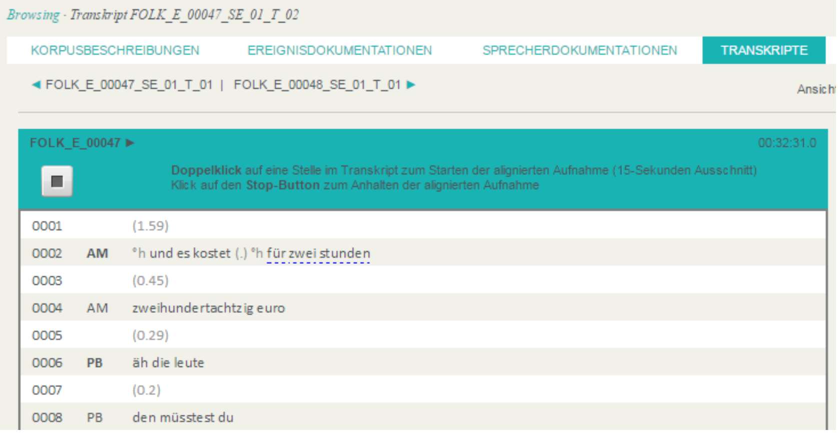
The browsing mode of the DGD gives access to individual data sets in FOLK. Starting from a tabular overview (see figure 19), users can navigate and view metadata on speech events and speakers, listen to audio files and read transcripts.



#	Ereignis-ID ▲ ▼	Region ▲ ▼	Sprechereignis-Art ▲ ▼	Erhebungsdatum ▲ ▼
1	FOLK_E_00001 ▶	Rheinfränkische Sprachregion	Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	2009
2	FOLK_E_00002 ▶	Rheinfränkische Sprachregion	Alltagsgespräch: Vorlesen für Kinder	2009
3	FOLK_E_00003 ▶	Obersächsische Sprachregion	Institutionelle Kommunikation: Prüfungsgespräch in der Hochschule	2010
4	FOLK_E_00004 ▶	Rheinfränkische Sprachregion	Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	2009
5	FOLK_E_00005 ▶	Rheinfränkische Sprachregion	Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	2009
6	FOLK_E_00006 ▶	Rheinfränkische Sprachregion	Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	2009
7	FOLK_E_00007 ▶	Rheinfränkische Sprachregion	Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	2009
8	FOLK_E_00008 ▶	Rheinfränkische Sprachregion	Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	2009
9	FOLK_E_00009 ▶	Rheinfränkische Sprachregion	Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	2009
10	FOLK_E_00010 ▶	Bairische Sprachregion	Alltagsgespräch: Spielinteraktion mit Kindern	2009

Figure 19: Tabular overview of FOLK speech events

Hyperlinks between the representations of the different data types allow for an exploration of the relationships between them. For instance, starting from the metadata for a given speaker, the speech event(s) this speaker participates in can be displayed, and from there, a link to the corresponding audio file(s) and transcript(s) is available. The transcript, in turn, contains links to the speech event and speaker metadata, and clicking on any word in the transcript will start playback of the corresponding part of the aligned audio (see figure 20).

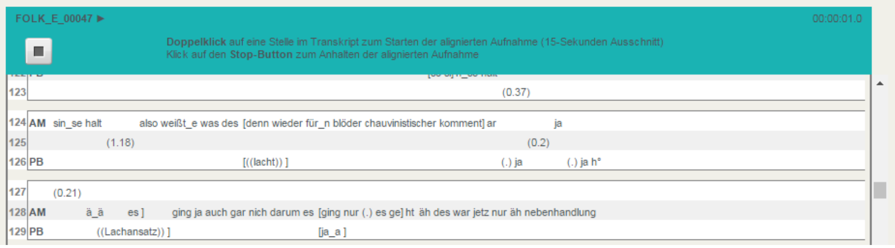


Doppelklick auf eine Stelle im Transkript zum Starten der alignierten Aufnahme (15-Sekunden Ausschnitt)
Klick auf den **Stop-Button** zum Anhalten der alignierten Aufnahme

0001		(1.59)
0002	**AM**	°h und es kostet (.) °h für zwei stunden
0003		(0.45)
0004	AM	zweihundertachtzig euro
0005		(0.29)
0006	**PB**	äh die leute
0007		(0.2)
0008	PB	den müsstest du

Figure 20: Display of a transcript with aligned audio (current playback position indicated by the dashed line)

The default display shows the transcript text in modified orthography together with non-speech tokens (pauses etc.) in a line-for-line notation (one contribution per line). Alternatively, FOLK transcripts can be displayed as musical scores (see figure 21), which makes it easier to understand the temporal flow of events (especially simultaneous and overlapping speech), or in a “normalised” version which displays the text in standard orthography and omits all non-speech tokens, making it easier to read for users who are not familiar with the specialised transcription forms.

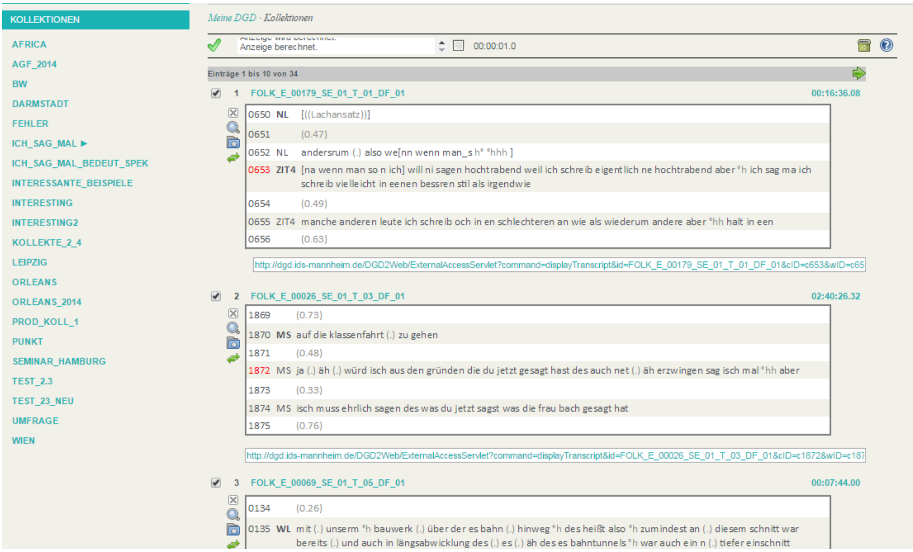


Doppelklick auf eine Stelle im Transkript zum Starten der alignierten Aufnahme (15-Sekunden Ausschnitt)
Klick auf den **Stop-Button** zum Anhalten der alignierten Aufnahme

```
123                                                                   (0.37)

124 AM  sin_se halt      also weißt_e was des [denn wieder für_n blöder chauvinistischer komment] ar       ja
125            (1.18)                                                        (0.2)
126 PB                                [((lacht)) ]                      (.) ja       (.) ja h°

127     (0.21)
128 AM        ä_ä     es ]     ging ja auch gar nich darum es [ging nur (.) es ge] ht äh des war jetz nur äh nebenhandlung
129 PB           ((Lachansatz)) ]                      [ja_a ]
```

Figure 21: Transcript in musical score display

In order to retain relevant and interesting excerpts that are identified in the browsing process, users can add them to a collection and store them inside the database (figure 22).



Anzeige berechnet.

Einträge 1 bis 10 von 34

☑ 1 FOLK_E_00179_SE_01_T_01_DF_01 00:16:36.08

```
0650 NL   [((Lachansatz))]
0651      (0.47)
0652 NL   andersrum (.) also we[nn wenn man_s h° °hhh ]
0653 ZIT4 [na wenn man so n ich] will ni sagen hochtrabend weil ich schreib eigentlich ne hochtrabend aber °h ich sag ma ich
          schreib vielleicht in eenen bessren stil als irgendwie
0654      (0.49)
0655 ZIT4 manche anderen leute ich schreib och in en schlechteren an wie als wiederum andere aber °hh halt in een
0656      (0.63)
```

http://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00179_SE_01_T_01_DF_01&cID=c653&wID=c65

☑ 2 FOLK_E_00026_SE_01_T_03_DF_01 02:40:26.32

```
1869    (0.73)
1870 MS auf die klassenfahrt (.) zu gehen
1871    (0.48)
1872 MS ja (.) äh (.) würd isch aus den gründen die du jetzt gesagt hast des auch net (.) äh erzwingen sag isch mal °hh aber
1873    (0.33)
1874 MS isch muss ehrlich sagen des was du jetzt sagst was die frau bach gesagt hat
1875    (0.76)
```

http://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00026_SE_01_T_03_DF_01&cID=c1872&wID=c187

☑ 3 FOLK_E_00069_SE_01_T_05_DF_01 00:07:44.00

```
0134    (0.26)
0135 WL mit (.) unserm °h bauwerk (.) über der es bahn (.) hinweg °h des heißt also °h zumindest an (.) diesem schnitt war
        bereits (.) und auch in längsabwicklung des (.) es (.) äh des es bahntunnels °h war auch ein n (.) tiefer einschnitt
```

Figure 22: A user-generated collection of transcript excerpts

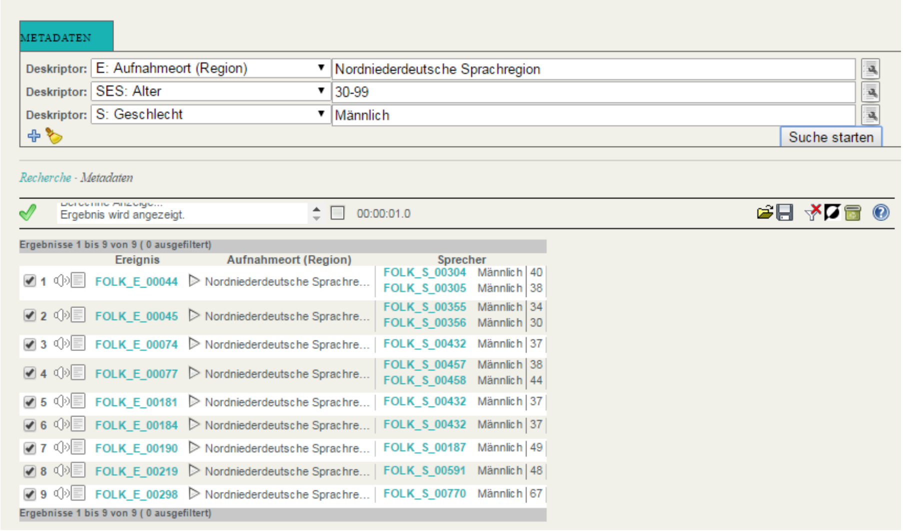
3.2 Query

Queries on DGD data can be carried out in three different manners.

First, a **full text search** is a simple way of searching through metadata and transcripts to obtain a global impression of the occurrences of a given term. Being realised via Oracle's full text functionality, the full text search is both relatively quick and flexible in so far as it does not make advanced assumptions about document structures. It can thus be used on different types of XML documents (such as metadata XML and FOLKER's transcript XML as illustrated above) and on plain text or PDF files (these are both file types in which the transcripts of some legacy corpora in the DGD are stored).

Since full text search, however, by definition, gets rid of most of the structure represented in XML elements and attributes, it is not suitable to exploit the FOLK data in its full complexity, including the different annotation levels and the links between transcripts and metadata. The DGD therefore also offers two types of 'structured' searches.

The **structured metadata query** is a means of finding speech events or speakers with certain properties. For instance, a structured metadata search could be used on FOLK to find all instances of interactions with male participants older than 30 years from Northern Germany (as in figure 23). The result of such a query is a list of matching speech events, possibly combined with a list of matching speakers. The list can be saved as a virtual corpus and used as a basis for structured token searches.



Deskriptor: E: Aufnahmeort (Region) ▾ Nordniederdeutsche Sprachregion

Deskriptor: SES: Alter ▾ 30-99

Deskriptor: S: Geschlecht ▾ Männlich

✓ Berechne Anzeige...
Ergebnis wird angezeigt. □ 00:00:01.0

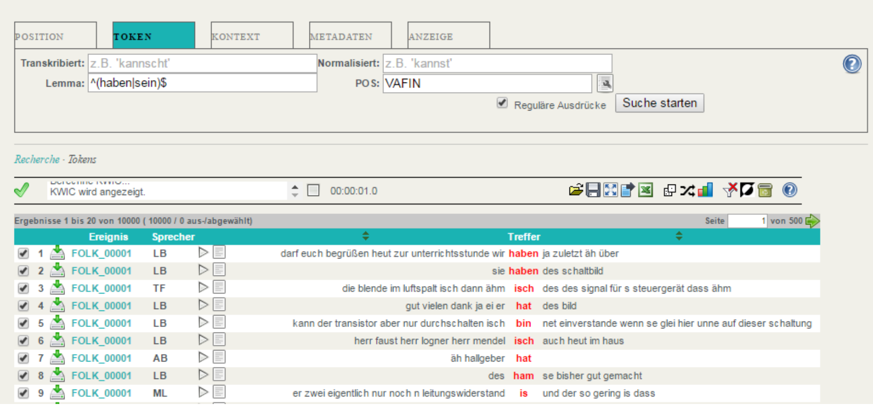
Ergebnisse 1 bis 9 von 9 (0 ausgefiltert)

		Ereignis	Aufnahmeort (Region)	Sprecher		
☑	1	FOLK_E_00044	▷ Nordniederdeutsche Sprachre...	FOLK_S_00304 FOLK_S_00305	Männlich Männlich	40 38
☑	2	FOLK_E_00045	▷ Nordniederdeutsche Sprachre...	FOLK_S_00355 FOLK_S_00356	Männlich Männlich	34 30
☑	3	FOLK_E_00074	▷ Nordniederdeutsche Sprachre...	FOLK_S_00432	Männlich	37
☑	4	FOLK_E_00077	▷ Nordniederdeutsche Sprachre...	FOLK_S_00457 FOLK_S_00458	Männlich Männlich	38 44
☑	5	FOLK_E_00181	▷ Nordniederdeutsche Sprachre...	FOLK_S_00432	Männlich	37
☑	6	FOLK_E_00184	▷ Nordniederdeutsche Sprachre...	FOLK_S_00432	Männlich	37
☑	7	FOLK_E_00190	▷ Nordniederdeutsche Sprachre...	FOLK_S_00187	Männlich	49
☑	8	FOLK_E_00219	▷ Nordniederdeutsche Sprachre...	FOLK_S_00591	Männlich	48
☑	9	FOLK_E_00298	▷ Nordniederdeutsche Sprachre...	FOLK_S_00770	Männlich	67

Ergebnisse 1 bis 9 von 9 (0 ausgefiltert)

Figure 23: Metadata query on FOLK resulting in a virtual corpus

The **structured token query**, finally, is the core component of the DGD. Its base functionality is to allow the user to specify one or more properties of a token, such as its transcribed or orthographic form, its lemma and/or its part of speech, and to display as a KWIC concordance all the matching tokens in the selected corpus or corpora (see figure 24). Properties can be specified as plain strings or as string patterns in the form of regular expressions.



Transkribiert: z.B. 'kannscht'
Normalisiert: z.B. 'kannst'
Lemma: ^(haben|sein)$
POS: VAFIN
☑ Reguläre Ausdrücke

KWIC wird angezeigt. 00:00:01.0

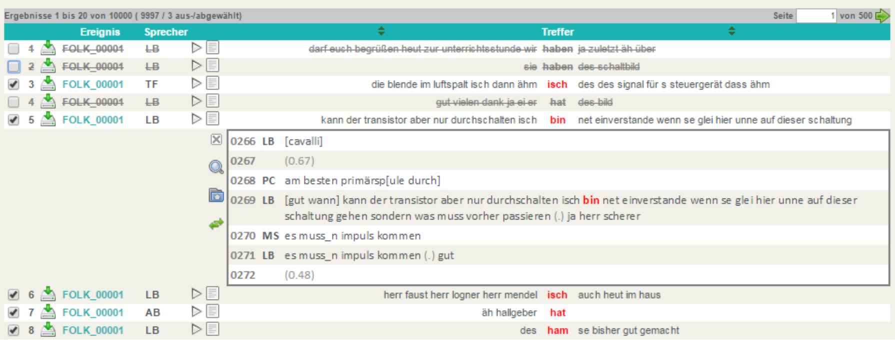
Ergebnisse 1 bis 20 von 10000 (10000 / 0 aus-/abgewählt) Seite 1 von 500

		Ereignis	Sprecher	Treffer		
☑	1	FOLK_00001	LB	darf euch begrüßen heut zur unterrichtsstunde wir	haben	ja zuletzt äh über
☑	2	FOLK_00001	LB	sie	haben	des schaltbild
☑	3	FOLK_00001	TF	die blende im luftspalt isch dann ähm	isch	des des signal für s steuergerät dass ähm
☑	4	FOLK_00001	LB	gut vielen dank ja ei er	hat	des bild
☑	5	FOLK_00001	LB	kann der transistor aber nur durchschalten isch	bin	net einverstande wenn se glei hier unne auf dieser schaltung
☑	6	FOLK_00001	LB	herr faust herr logner herr mendel	isch	auch heut im haus
☑	7	FOLK_00001	AB	äh hallgeber	hat	
☑	8	FOLK_00001	LB	des	ham	se bisher gut gemacht
☑	9	FOLK_00001	ML	er zwei eigentlich nur noch n leitungswiderstand	is	und der so gering is dass

Figure 24: Token Query for lemmas ‚haben' or 'sein' as finite auxiliaries (POS=VAFIN)

A query will thus always start with a concordance for a single class of tokens. Additional functionality allows the user to further explore and refine this result.

The refinement can be done manually by inspecting individual search results. Audio playback for the corresponding part of the recordings is available directly from the KWIC concordance. For each line of the concordance, metadata about the corresponding speech event and speaker can be displayed. To explore the interaction context of the matching token, the corresponding transcript excerpt can be folded out underneath the KWIC line (see figure 25). In that way, automated query can be combined with detailed qualitative analyses. By deselecting individual lines of the concordance, users have the possibility to clean the result from false positives identified in that process.



Ergebnisse 1 bis 20 von 10000 (9997 / 3 aus-/abgewählt) Seite 1 von 500

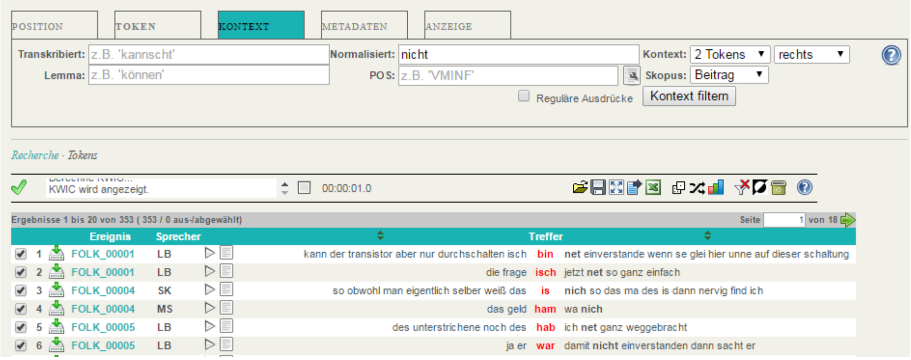
		Ereignis	Sprecher	Treffer		
□	~~1~~	~~FOLK_00001~~	~~LB~~	~~darf euch begrüßen heut zur unterrichtsstunde wir~~	~~haben~~	~~ja zuletzt äh über~~
□	~~2~~	~~FOLK_00001~~	~~LB~~	~~sie~~	~~haben~~	~~des schaltbild~~
☑	3	FOLK_00001	TF	die blende im luftspalt isch dann ähm	isch	des des signal für s steuergerät dass ähm
□	~~4~~	~~FOLK_00001~~	~~LB~~	~~gut vielen dank ja ei er~~	~~hat~~	~~des bild~~
☑	5	FOLK_00001	LB	kann der transistor aber nur durchschalten isch	bin	net einverstande wenn se glei hier unne auf dieser schaltung

```
0266 LB  [cavalli]
0267     (0.67)
0268 PC  am besten primärsp[ule durch]
0269 LB  [gut wann] kann der transistor aber nur durchschalten isch bin net einverstande wenn se glei hier unne auf dieser
         schaltung gehen sondern was muss vorher passieren (.) ja herr scherer
0270 MS  es muss_n impuls kommen
0271 LB  es muss_n impuls kommen (.) gut
0272     (0.48)
```

		Ereignis	Sprecher	Treffer		
☑	6	FOLK_00001	LB	herr faust herr logner herr mendel	isch	auch heut im haus
☑	7	FOLK_00001	AB	äh hallgeber	hat	
☑	8	FOLK_00001	LB	des	ham	se bisher gut gemacht

Figure 25: KWIC with deselected lines (1,2 and 4) and transcript folded out (line 5)

It is also possible to use additional filters on a search result. Via the ‘Context’ tab, the properties of further tokens in the context of the matching token can be specified. ‘Context’ here can be limited to single contributions (the default case), to all contributions of the respective speaker or to the entire transcript. The context window can be specified as left, right or both-sided and in terms of token distance (see figure 26).



Transkribiert: z.B. 'kannscht'
Lemma: z.B. 'können'
Normalisiert: nicht
POS: z.B. 'VMINF'
Kontext: 2 Tokens ▾ rechts ▾
Skopus: Beitrag ▾
□ Reguläre Ausdrücke

KWIC wird angezeigt. □ 00:00:01.0

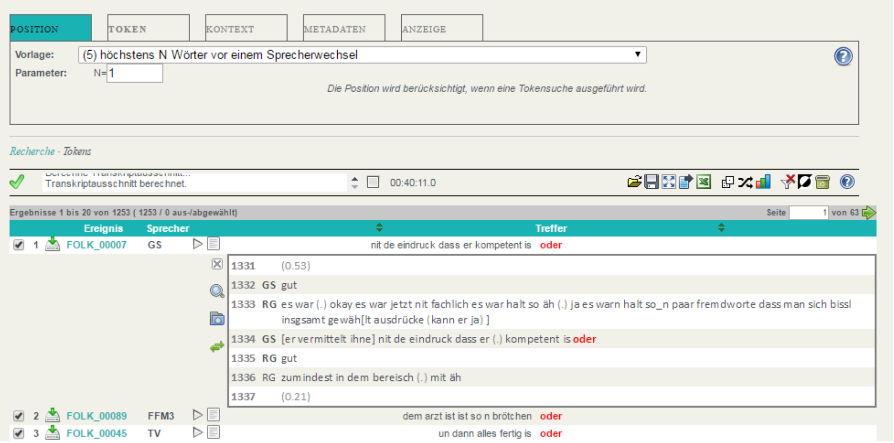
Ergebnisse 1 bis 20 von 353 (353 / 0 aus-/abgewählt) Seite 1 von 18

		Ereignis	Sprecher		Treffer		
☑	1	FOLK_00001	LB		kann der transistor aber nur durchschalten isch	bin	net einverstande wenn se glei hier unne auf dieser schaltung
☑	2	FOLK_00001	LB		die frage	isch	jetzt net so ganz einfach
☑	3	FOLK_00004	SK		so obwohl man eigentlich selber weiß das	is	nich so das ma des is dann nervig find ich
☑	4	FOLK_00004	MS		das geld	ham	wa nich
☑	5	FOLK_00005	LB		des unterstrichene noch des	hab	ich net ganz weggebracht
☑	6	FOLK_00005	LB		ja er	war	damit nicht einverstanden dann sacht er

Figure 26: A search result filtered for the normalized form ‘nicht’ in a distance of two tokens in the right context

Using a context filter (also repeatedly, i.e. filtering first for one, then for another item in the context, which corresponds to a Boolean *and*) can thus serve to identify co-occurrences of two or more tokens. Items which do not match the filter will not be deleted immediately, but only deselected in the concordance. In that way, the effect of a filter can be evaluated (and, if necessary, reversed) in a transparent manner. Similarly, the ‘Metadata’ tab can be used to filter search results according to properties of the respective speech events or speakers. In an analogous manner to the structured metadata query, users can, for instance, specify a metadata filter for conversations including male speakers from a certain region.

A filter type specific to interaction data is implemented in the “Position” tab where the user can (before carrying out the actual token query) restrict searches to specific positions in the interaction, such as “within *n* tokens of the beginning/end of a contribution” / “within *n* tokens of a change of speaker” / “inside or immediately before/after an overlap” / “in the vicinity of a pause” (see figure 27). Making queries sensitive to the interactive structure of the FOLK data is especially useful for investigating phenomena which conversation analysis and related fields are interested in. In particular, it enables the study of functional aspects of certain items (such as discourse markers) in speakers’ organisation of turn-taking. An example would be corpus-guided studies of turn-initiations as described in Heritage (2013).



Vorage: (5) höchstens N Wörter vor einem Sprecherwechsel

Parameter: N=1

Die Position wird berücksichtigt, wenn eine Tokensuche ausgeführt wird.

Transkriptausschnitt berechnet.

00:40:11.0

Ergebnisse 1 bis 20 von 1253 (1253 / 0 aus-/abgewählt)

Seite 1 von 63

		Ereignis	Sprecher		Treffer
☑	1	FOLK_00007	GS	▷	nit de eindruck dass er kompetent is **oder**
☑	2	FOLK_00089	FFM3	▷	dem arzt ist ist so n brötchen **oder**
☑	3	FOLK_00045	TV	▷	un dann alles fertig is **oder**

```
1331      (0.53)
1332  GS  gut
1333  RG  es war (.) okay es war jetzt nit fachlich es war halt so äh (.) ja es warn halt so_n paar fremdworte dass man sich bissl
          insgsamt gewäh[lt ausdrücke (kann er ja) ]
1334  GS  [er vermittelt ihne] nit de eindruck dass er (.) kompetent is oder
1335  RG  gut
1336  RG  zumindest in dem bereisch (.) mit äh
1337      (0.21)
```

Figure 27: A query for the form ‚oder‘ restricted to the position immediately before a speaker change

Further operations can be carried out on the KWIC:

1 A KWIC can be stored in the database for later reuse, i.e. so that the underlying query and subsequent manual refinements do not have to be repeated.

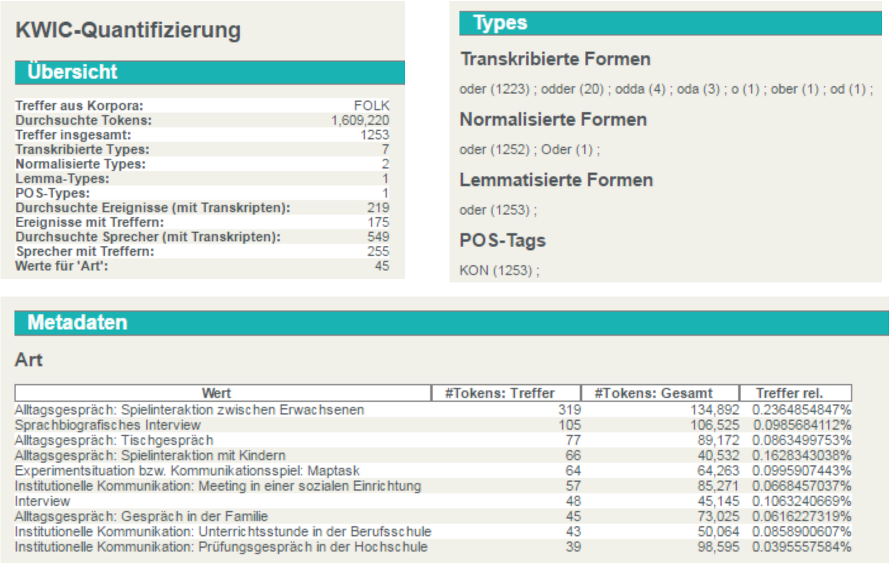
2 KWICs can be printed or exported to text or XML files. We notice that users especially value the possibility to export the KWIC in a file that can be further processed by spreadsheet applications such as MS Excel.

3 A random sample of an arbitrary size can be extracted from a KWIC. This is useful for obtaining a non-biased excerpt of a large result in order to keep manual inspection manageable.

4 KWICs can be scrambled randomly.

5 KWICs can be sorted according to any of the available columns. When sorting is applied to the left or right context column, it can serve to visually identify prominent co-occurrence patterns.

6 KWICs can be quantified, giving a concise summary of the number of tokens and types, of their combination with selected metadata parameters and of frequencies normalised with respect to the amount of data available (see figure 28).



KWIC-Quantifizierung

Übersicht

Treffer aus Korpora:	FOLK
Durchsuchte Tokens:	1,609,220
Treffer insgesamt:	1253
Transkribierte Types:	7
Normalisierte Types:	2
Lemma-Types:	1
POS-Types:	1
Durchsuchte Ereignisse (mit Transkripten):	219
Ereignisse mit Treffern:	175
Durchsuchte Sprecher (mit Transkripten):	549
Sprecher mit Treffern:	255
Werte für 'Art':	45

Types

Transkribierte Formen

oder (1223) ; odder (20) ; odda (4) ; oda (3) ; o (1) ; ober (1) ; od (1) ;

Normalisierte Formen

oder (1252) ; Oder (1) ;

Lemmatisierte Formen

oder (1253) ;

POS-Tags

KON (1253) ;

Metadaten

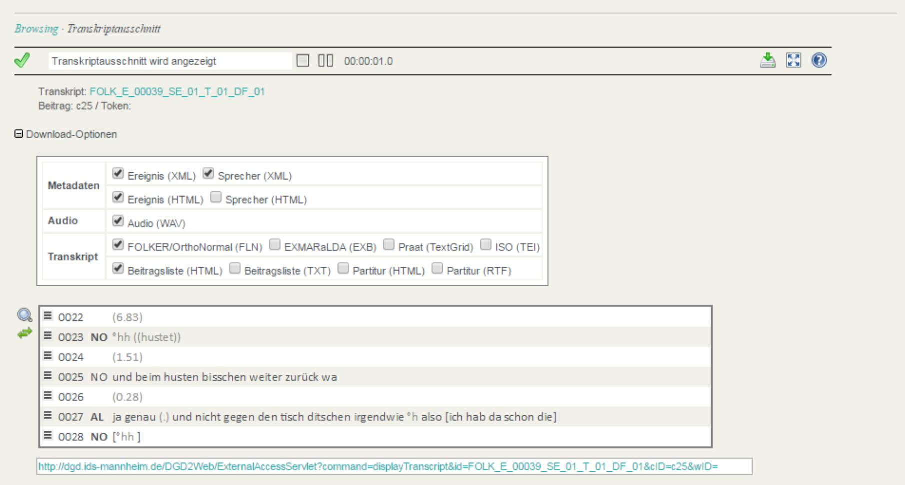
Art

Wert	#Tokens: Treffer	#Tokens: Gesamt	Treffer rel.
Alltagsgespräch: Spielinteraktion zwischen Erwachsenen	319	134,892	0.2364854847%
Sprachbiografisches Interview	105	106,525	0.0985684112%
Alltagsgespräch: Tischgespräch	77	89,172	0.0863499753%
Alltagsgespräch: Spielinteraktion mit Kindern	66	40,532	0.1628343038%
Experimentsituation bzw. Kommunikationsspiel: Maptask	64	64,263	0.0995907443%
Institutionelle Kommunikation: Meeting in einer sozialen Einrichtung	57	85,271	0.0668457037%
Interview	48	45,145	0.1063240669%
Alltagsgespräch: Gespräch in der Familie	45	73,025	0.0616227319%
Institutionelle Kommunikation: Unterrichtsstunde in der Berufsschule	43	50,064	0.0858900607%
Institutionelle Kommunikation: Prüfungsgespräch in der Hochschule	39	98,595	0.0395557584%

Figure 28: Various types of quantification for a search result: general figures (top left), types and tokens (top right), result counts relative to the metadata attribute ‘interaction type’ (bottom)

3.3 Download and Citation

The DGD’s browsing and query functionality as described in the two previous section addresses most requirements concerning the discovery of data relevant for a given analysis purpose – if it is in the data, the DGD user has a good chance of finding it there. However, FOLK (and, in fact, oral corpora in general) captures in its transcriptions and annotations only a selected part of the phenomena audible in the recordings, and quite a few phenomena that can play a role for linguistic analysis are thus not available for complete analysis inside the DGD. Prominent cases in point are prosodic features of speech like intonation and stress which are not taken into account in FOLK’s minimal transcriptions. An essential feature of the DGD is therefore that it enables the user to go beyond the information captured in the existing transcriptions and annotations and beyond the analysis functions offered by the web platform by downloading data for relevant excerpts onto a local computer and further processing them there.



Browsing · *Transkriptausschnitt*

Transkriptausschnitt wird angezeigt 00:00:01.0

Transkript: FOLK_E_00039_SE_01_T_01_DF_01
Beitrag: c25 / Token:

⊟ Download-Optionen

Metadaten	☑ Ereignis (XML) ☑ Sprecher (XML)
	☑ Ereignis (HTML) ☐ Sprecher (HTML)
Audio	☑ Audio (WAV)
Transkript	☑ FOLKER/OrthoNormal (FLN) ☐ EXMARaLDA (EXB) ☐ Praat (TextGrid) ☐ ISO (TEI)
	☑ Beitragsliste (HTML) ☐ Beitragsliste (TXT) ☐ Partitur (HTML) ☐ Partitur (RTF)

```
0022      (6.83)
0023 NO   °hh ((hustet))
0024      (1.51)
0025 NO   und beim husten bisschen weiter zurück wa
0026      (0.28)
0027 AL   ja genau (.) und nicht gegen den tisch ditschen irgendwie °h also [ich hab da schon die]
0028 NO   [°hh ]
```

http://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet?command=displayTranscript&id=FOLK_E_00039_SE_01_T_01_DF_01&cID=c25&wID=

Figure 29: Download of different data types and formats for a transcript excerpt

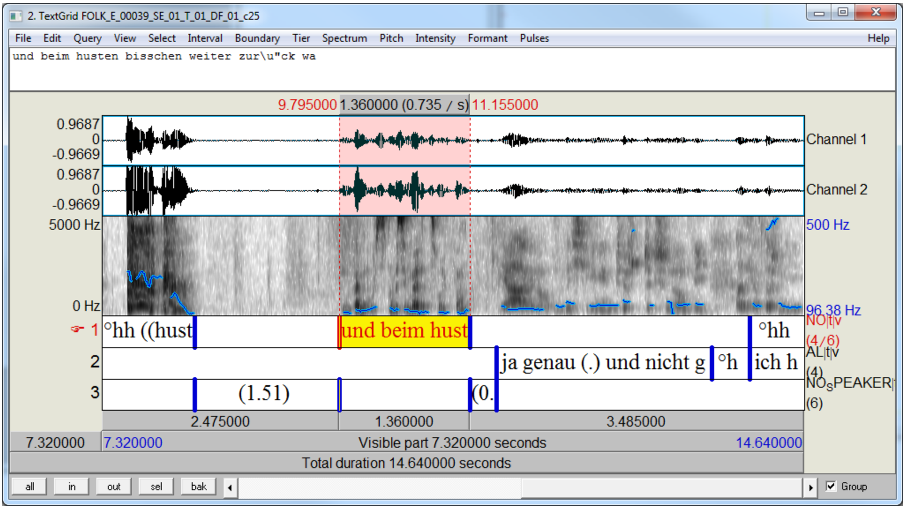
Different formats are offered for all data types. Metadata can be downloaded as XML or HTML, audio as PCM-WAV. Visualisations of transcripts are available, for instance, as RTF files for integration into text processing software, and, most importantly, the transcripts themselves can be downloaded:

1 as FOLKER XML files to be further processed (e.g. segmented, retranscribed) with the FOLKER or OrthoNormal tools described above, or

2 as EXMARaLDA XML files to be further processed (e.g. annotated on additional tiers) with the EXMARaLDA system, or

3 as Praat TextGrid to be further processed (e.g. subjected to instrumental phonetic analysis) with the Praat software (see figure 30), or

4 as TEI XML conforming to the guidelines of the Text Encoding Initiative and the ISO standard "ISO 24624:2016: Language resource management -- Transcription of spoken language" to open further possibilities of interoperating with other tools.



```
und beim husten bisschen weiter zur\u"ck wa
```

Figure 30: Audio and transcript excerpt with different visualisations of the audio signal in Praat

Users of the data are thus enabled to re-enter the workflow depicted in figure 1 at the stages of transcription or annotation, and the decisions the FOLK project makes in order to reduce the annotation effort do not have to result in a principle obstacle for certain types of exploitation of the audio data.

Finally, the DGD also makes it possible to directly address transcript excerpts via a single URL. This is meant to support citations of data, for instance in publications, where readers may want to not just read the transcript, but also get back to the audio. As an example, consider the following link which takes a registered user to the transcript excerpt depicted in figure 29:

http://dgd.ids-mannheim.de/DGD2Web/ExternalAccessServlet
?command=displayTranscript&id=FOLK_E_00039_SE_01_T_01_DF_01&cID=c25&wID=w41

4. Outlooks

The FOLK corpus itself as well as the tools and platform described here are under active development. Seven years after the start of the FOLK project, six years after the first full versions of the transcription and annotation tools and four years after the launch of the beta version of the DGD, we are confident that our workflow for corpus construction and dissemination is effective for our purposes and need not change in any fundamental way in the near future.

Regarding the corpus construction tools, this is corroborated also by the fact that other projects have made productive use of the transcription and annotation tools. This includes projects independent of FOLK, such as the Hamburg Corpus of Bilingual Language Acquisition (HABLA, Kupisch et al. 2012) or the Last Minute Corpus (Rösner et al. 2012), as well as projects with which we have been or are in a close collaboration, such as different ongoing research and dissertation projects. Typically, in such collaborations, researchers receive technical advice from us in return for a part of their transcribed data which we integrate into FOLK. By making the tools themselves freely available, adequately documenting their use (through manuals and guidelines) and offering training and support also for external researchers, we hope, in the long run, to be able to motivate more and more colleagues for this kind of joint effort. Ideally, the practice surrounding the tools for corpus compilation will thus have a measurable effect on the development of the corpus itself, and future extensions of FOLK will profit more and more from external contributions.

The corpus construction workflow has developed gradually, combining existing tools and methods (like EXMARaLDA and GAT) wherever possible, extending and adapting them (to FOLKER and cGAT, respectively, for instance) wherever necessary. In the long run, an obvious option for improvement lies in a tighter integration of the individual components, for instance a more direct interfacing of the annotation tools with the instruments for managing metadata and corpus organisation. Ideally, and following a general trend in this area (see, for example, tools like WebAnno in CLARIN, de Castilho et al. 2014), the workflow could be remodelled in an integrated web-based environment, meaning that "manual" standalone tools for annotation like FOLKER and OrthoNormal would become browser applications in a client-server architecture, and that automatic processes like orthographic normalisation and POS tagging would be realised as web-services. Such an environment would also make it easier to distribute tasks to external partners and, ultimately, make a modest form of "crowd sourcing" a realistic option for FOLK. First steps towards implementing components of the workflow as web-services have already been taken (see Schmidt et al. 2017). The crowd-sourcing aspect for transcriptions is currently explored in a pilot project at the AGD.

The DGD as the corpus dissemination tool is used not only for FOLK, but also for most other spoken language corpora at the Archive for Spoken German (such as large dialect corpora, see Stift/Schmidt 2014). The transcriptions of these corpora can be accommodated by FOLKER's data model, and typically have a somewhat simpler structure: most of them are transcribed orthographically, so there is no need for a second normalisation layer, non-speech tokens like pauses and non-verbal descriptions play a less prominent role, and the interaction structure is often not represented in as much detail (especially regarding overlaps). The browsing and query mechanisms suitable for FOLK are therefore usually more than sufficient also for this other type of data.

Besides the obligatory maintenance requirements and future extensions of FOLK and other corpora in the DGD, we see three areas as prioritized for the development of new functionality in the platform:

1 Since it is becoming more and more common to study spoken language on the basis of video data – making it possible to take into account also the embodied dimension of interaction –, the DGD, as a minimum requirement, will have to provide means of accessing videos in its browsing and query modes. Roughly a third of the FOLK data are already available as digital video files, and we plan to integrate these data alongside suitable visualisation methods into the DGD in the near future.

2 An obvious user need when working with the DGD is to save and retrieve virtual corpora, collections and search results not just for individual use, but also for collaborative

work involving other users. We are therefore working on mechanisms of sharing such data in personal and group workspaces inside the platform.

3 So far, the platform is ready to exploit annotations only on the token level, which is the only type of annotation so far included in FOLK and in all other corpora of the Archive for Spoken Language. There are many cases, however, where spoken language transcriptions are annotated on larger segments, for example for pragmatic functions of chunks or utterances or for conversation topics of larger stretches of a transcript. As we can see already on the occasion of the planned integration of a new resource in the DGD – the GeWiss corpus of Academic Speech (Fandrych et al. 2012), which has been partly annotated for discourse comments as well as for quotes and references – such annotations will have to be accommodated by the data model as well as made accessible through the browsing and query interfaces.

In principle, we think that the DGD interface could also be usable and useful for spoken language corpora constructed or archived in other contexts. Several such resources have become available in the last years, such as the ESLO corpus (Eshkol-Taravella 2012) and corpora in the CLAPI database (Groupe ICOR, in press) for French, the oral parts of the Czech National corpus (Kren 2015) or the Slovene GOS corpus (Verdonik et al. 2013), to name just a few. However, in contrast to the situation for annotation tools, where long-standing development efforts combined with interoperability improvements (see for example Schmidt et al. 2008) have led to a fair degree of conversion, we find that Anthony's (2009) observation that "[Tools widely used by corpus linguists] all offer a different user-experience, because each tool is created in isolation and thus offers a different user interface, control flow, and functionality" is still largely true for tools providing access to spoken language corpora. Technically, the DGD is not ready to be transferred to other contexts, but we hope that, in the mid-term, its design can serve as one source of inspiration for an effort to develop tools for accessing spoken language corpora that are less bound to a specific institutional context.

References

ANTHONY, L. (2009). „Issues in the design and development of software tools for corpus studies: The case for collaboration." In: Baker, P. (ed.), Contemporary corpus linguistics. London: Continuum Press, pp. 87-104.

ECKART DE CASTILHO, R. (no date). „TreeTagger for Java – TT4J". [https://reckart.github.io/tt4j/]

ECKART DE CASTILHO, R. / BIEMANN, C. / GUREVYCH, I. / YIMAM, S.M. (2014). „WebAnno: a flexible, web-based annotation tool for CLARIN". In: Proceedings of the CLARIN Annual Conference (CAC) 2014, Soesterberg, Netherlands. Linköping University Electronic Conference Proceedings

ESHKOL-TARAVELLA, I. / BAUDE, O. / MAUREL, D. / HRIBA, L. / DUGUA, C. / TELLIER, I., (2012). „Un grand corpus oral ‚disponible' : le corpus d'Orléans 1968-2012." In: Ressources linguistiques libres, TAL. 52,3/2011, pp. 17-46.

FANDRYCH, C. / MEIßNER, C. / SLAVCHEVA, A. (2012). „The GeWiss Corpus: Comparing Spoken Academic German, English and Polish." In: Schmidt, T., Wörner, K. (eds.): Multilingual Corpora and Multilingual Corpus Analysis. Hamburg Studies in Multilingualism (14). Amsterdam: Benjamins, pp. 319-337.

FANDRYCH, C. / FRICK, E. / HEDELAND, H. / ILIASH, A. / JETTKA, D. / MEIßNER, C. / SCHMIDT, T. / WALLNER, F. / WEIGERT, K. / WESTPFAHL, S. (2016). „User, who art thou? User Profiling for Oral Corpus Platforms. " In: Proceedings of the 10th Conference on International Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 280-287. [http://nbn-resolving.de/urn:nbn:de:bsz:mh39-50774]

FIEHLER, R. / WAGENER, P. (2005). „Die Datenbank Gesprochenes Deutsch (DGD) – Sammlung, Archivierung und Untersuchung gesprochener Sprache als Aufgaben der Sprachwissenschaft." In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion. 6/2005, pp. 136-147. [http://nbn-resolving.de/urn:nbn:de:bsz:mh39-6869]

GASCH, J. / BRINCKMANN, C. / DICKGIEßER, S. (2008). „memasysco: XML schema based metadata management system for speech corpora." In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesch, Marokko. Paris: European Language Resources Association (ELRA), pp. 2865-2870 [http://www.lrec-conf.org/proceedings/lrec2008/pdf/729_paper.pdf]

GROUPE ICOR (H. BALDAUF-QUILLIATRE, I. COLON DE CARVAJAL, C. ETIENNE, E. JOUIN-CHARDON, S. TESTON-BONNARD, V. TRAVERSO) (IN PRESS). „CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes." In Avanzi, M., Béguelin, M.-J. & Diémoz, F. (eds), Corpus de français parlés et français parlés des corpus, Cahiers Corpus.

HERITAGE, J. (2013). „Turn-initial position and some of its occupants." Journal of Pragmatics (2013), [http://dx.doi.org/10.1016/j.pragma.2013.08.025]

HUNSTON, S. (2002). „Corpora in applied linguistics." Cambridge: Cambridge University Press.

INSTITUT FÜR DEUTSCHE SPRACHE (2014). „Korpusbasierte Wortformenliste DeReWo.", Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland. [http://www.ids-mannheim.de/derewo]

KŘEN, M. (2015). „Recent Developments in the Czech National Corpus." In: Bański, P., Biber, H., Breiteneder, E., Kupietz, M., Lüngen, H., Witt, A. (eds.) (2015): Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3). Mannheim: Institut für Deutsche Sprache, pp. 1-4.

KUPIETZ, M. / SCHMIDT, T. (2015). „Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung." In: Eichinger, L. M. (ed.): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. Berlin/Boston: de Gruyter. (Jahrbuch des Instituts für Deutsche Sprache 2014), pp. 297-322. [http://nbn-resolving.de/urn:nbn:de:bsz:mh39-34824]

KUPISCH, T. / BARTON, D. / BIANCHI, G. / STANGEN, I. (2012). „The HABLA-corpus (German-French and German-Italian)." In: Schmidt, T. & Wörner, K. (eds.): Multilingual Corpora and Multilingual Corpus Analysis. Amsterdam: Benjamins, pp. 63-179.

RÖSNER, D. / FROMMER, J. / FRIESEN, R. / HAASE, M. / LANGE, J. / OTTO, M. (2012). „LAST MINUTE: A Multimodal Corpus of Speech-based User-Companion Interactions." In: : Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey. Paris: European Language Resources Association (ELRA), pp. 2559-2566 [http://www.lrec-conf.org/proceedings/lrec2012/pdf/550_Paper.pdf]

SCHMID, H. (1994). „Probabilistic Part-of-Speech Tagging Using Decision Trees." Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

SCHMIDT, T. (2016). „Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German". In: Kirk, J. M. and Andersen, G. (eds.): Compilation, transcription, markup and annotation of spoken corpora, Special Issue of the International Journal of Corpus Linguistics [IJCL 21:3], pp. 396-418. [http://dx.doi.org/10.1075/ijcl.21.3.05sch]

SCHMIDT, T. (2014). „The Database for Spoken German – DGD2.” In: Proceedings of the 9th Conference on International Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland. Paris: European Language Resources Association (ELRA), pp. 1451-1457. [http://nbn-resolving.de/urn:nbn:de:bsz:mh39-24425]

SCHMIDT, T./DUNCAN, S. / EHMER, O. / HOYT, J. / KIPP, M. / LOEHR, D. / MAGNUSSON, M. / ROSE, T. / SLOETJES, H. (2009). „An exchange format for multimodal annotations.” In: Kipp, M., Martin, J.-C., Paggio, P., Heylen, D. (eds.): Multimodal corpora: from models of natural interaction to systems and applications. Berlin/Heidelberg: Springer, 2009, pp. 207-221.

SCHMIDT, T. / SCHÜTTE, W. (2010). „FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction.” In: Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta. Paris: European Language Resources Association (ELRA), pp. 2091-2096. [http://nbn-resolving.de/urn:nbn:de:bsz:mh39-22323]

SCHMIDT, T. / WÖRNER, K. (2014). „EXMARaLDA.” In: Durand, J., Gut, U., Kristoffersen, G. (eds.): The Oxford Handbook of Corpus Phonology. Oxford: OUP 2014, pp. 402-419.

SCHMIDT, T. / HEDELAND, H. / JETTKA, D. (2017). „Conversion and Annotation Web Services for Spoken Language Data in CLARIN.” To appear in: Proceedings of the CLARIN Annual Conference (CAC) 2016, Aix en Provence, France. Linköping University Electronic Conference Proceedings

STIFT, U.-M. / SCHMIDT, T. (2014). „Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch.” In: Institut für Deutsche Sprache (Hrsg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Redaktion: Melanie Steine, Franz Josef Berens. S. 360-375 - Mannheim: Institut für Deutsche Sprache, 2014. [http://nbn-resolving.de/urn:nbn:de:bsz:mh39-24779]

WESTPFAHL, S. / SCHMIDT, T. (2016). „FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German.” In: Proceedings of the 10th Conference on International Language Resources and Evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 1493-1499. [http://nbn-resolving.de/urn:nbn:de:bsz:mh39-50786]

VERDONIK, D. / KOSEM, I. / ZWITTER-VITEZ, A. / KREK, S. / STABEJ, M. (2013). „Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS.” Language resources and evaluation, Dec. 2013, vol. 47, iss. 4, pp. 1031-1048. [http://dx.doi.org/10.1007/s10579-013-9216-5].