Anne Ferger, André Frank Krause, and Karola Pitsch

A workflow for creating, harmonizing and analyzing structured corpora of multimodal interaction

Abstract: Creating structured and consistent corpus resources for the analysis of multimodal interaction, computer-mediated communication, and socio-technical settings is generally a time-consuming and meticulous task. It involves dealing with various media formats (e.g., audiovisual, eye-tracking, log files), time alignment, and modelling multimodal aspects of communication, including many manual changes to these formats. Based on our work creating a structured corpus of human-robot interaction, we present a workflow focussing on automation, standard formats, and a sustainable approach to research data management. This workflow leverages recent developments in spoken language corpora and takes them beyond mere text and speech. It includes automated procedures to enrich data (e.g., part-of-speech tagging) to achieve higher data consistency and to convert data into a set of standard formats (e.g., TEI XML, dataFrame) from which calculations, visualizations, etc. can be generated for further analysis. Automating this workflow using git for version control and a GitLab Continuous Integration functionality, these procedures are reapplied whenever changes are made to the source data, so that amendments to transcripts, for example, can be reintroduced into the original transcript file. We show how higher data quality can be reached in these corpora and how the proposed workflow can be applied on corpora modelled following the CMC-core TEI schema. By exploring different analyses (including gaze, part-of-speech tagging, and time alignment) on the base of TEI XML documents originating from this workflow, we show how the resulting corpora offer more finely grained possibilities of analysis.

Keywords: multimodal interaction, socio-technical settings, corpus analysis, corpus workflow, TEI format

Anne Ferger, University of Duisburg-Essen, e-mail: anne.ferger@uni-due.de
André Frank Krause, Rhine-Waal University of Applied Sciences,
e-mail: andrefrank.krause@hochschule-rhein-waal.de
Karola Pitsch, University of Duisburg-Essen, e-mail: karola.pitsch@uni-due.de

https://doi.org/10.1515/9783111434018-009

1 Introduction

Different research communities develop resources for qualitative and quantitative analyses of multimodal interaction and socio-technical settings, including computer-mediated communication. For all of them, it is essential to generate multimodal corpora with high data quality and which can be analyzed from various disciplinary perspectives and with different research methods. In this vein, in the fields of interactional linguistics and conversation analysis, including the German *Gesprächsforschung*, approaches that link qualitative and quantitative methods have become more frequent as well (e.g., Pitsch et al. 2014; Stivers 2015; Kendrick and Holler 2017; Rühlemann 2018; Mundwiler et al. 2019; Luginbühl et al. 2021). They benefit from creating structured corpora (in the sense of Schmidt 2016) of interactional situations that are tailored both to human readability and technical means of analysis and thus enable the combination of qualitative and quantitative analyses in a dynamic way. Other applications to leverage these resources can be corpus queries in corpus linguistics (Schmidt 2016) and using the resources for second language teaching (Fandrych 2022).

Against this background, we address the following questions:

a) How can we best structure corpora of multimodal interaction which include novel types of time series data (e.g., robot log files, sensor output), which are human- and machine-readable, and model them using a standard format (e.g., TEI)?
b) How can we assure and enhance the quality of the corpus data with regard to inconsistencies, missing information, and enrichment?
c) How can these measures be applied continuously and automatically, and how can the workflows be reused on other (especially CMC) resources?
d) How can we prepare the corpus so that it allows for different export formats catering for different forms of storage and analyses?
e) How can the created structured corpora facilitate comprehensive analyses?

To address these questions, we suggest that it is beneficial to link tools and methods developed creating spoken language corpora (Schmidt 2016; Schmidt 2018; Hedeland and Ferger 2020; Arkhangelskiy, Hedeland, and Riaposov 2020; Ferger and Jettka 2021; Hirschmann and Schmidt 2022) with a research data management perspective (Hermann, Pietsch, and Cimiano 2021). In recent years within the field of research data management, a prominent focus has been given to the standardization and sustainability of research data formats. This is exemplified by projects like the German National Research Data Infrastructure in Germany (Kraft et al. 2021), for example. These developments are also essential for creating and processing structured corpora, as they lead to standardized corpus resources that are suitable for long-term archiving and are reusable in different contexts, as well as standardized methods, tools, and workflows that minimize efforts in future projects and increase the scientific reproducibility of analyses and their results.

In what follows, we present a semi-automatic workflow based on our experience working with multimodal and multisensorial data of human-robot interaction which includes – beyond common audiovisual data – log files from the robot's speech recognition and voice output in real time and sensor data from motion capture devices (e.g., Kinect). All data are synchronized via the timeline and share central features with chat exports of social media tools (e.g., from WhatsApp chats with time-synchronized text messages, audio recordings, or images).

2 Background: Workflows and tools for creating a structured corpus

For creating structured corpora of oral communication, ISO standard 24624:2016 "Language resource management – Transcription of spoken language" (Hedeland and Schmidt 2022) has been defined. It uses the framework format for encoding recommended by the Text Encoding Initiative (TEI Consortium 2023; the recommended format is abbreviated as TEI in the following). The TEI format is an XML-based format. Also, for the so-called task of "corpus compilation" (Schmidt 2016: 119), extensive workflows and tools exist. These were created when developing the FOLK corpus for spoken German language (Schmidt 2023), for example. Since the editor used to compile the FOLK corpus – called "FOLKER" (Schmidt and Schütte 2010) – is a timeline- and XML-based tool, procedures can be adapted for other data which are organized in timeline and XML format, such as data prepared with editors like EXMARaLDA (Schmidt and Wörner 2014) or ELAN (Sloetjes 2014), which are interoperable with regard to their data format. The TEI standard allows one to import data into the ZuMult tools (Fandrych et al. 2022) for analysis and querying of verbal data, but it does not include modelling of multimodal annotations or robot log files.

Some of these workflows and tools include methods for part of speech (POS) and lemma tagging (Westpfahl and Schmidt 2013; Westpfahl et al. 2017) using TreeTagger (Schmid 1995) (see chapter 4.3) and the use of the TEI format following ISO standard 24624:2016. In our workflow, we specifically adapted methods for converting ELAN files into TEI files. These methods included tokenizing verbal utterances based on methods used in the EXMARaLDA code and handling special cases such as robot log files and annotations of bodily conduct (see chapter 4.3). We also adapted and created new consistency checks (see chapter 4.2) and automated them using GitLab CI (see chapter 4.1).

In research on multimodal interaction and pragmatics, there are comprehensive works on corpus linguistics by Rühlemann, among others (Rühlemann 2017, 2018; Rühlemann and Gee 2017; Rühlemann and Ptak 2023). While these works propose XML formats for these corpora (e.g., Rühlemann 2017; Rühlemann and Gee 2017), the TEI format and its ISO standard for spoken language are not used in these cases, which creates barriers for sustainable, long-term archiving, machine readability, and reuse in other contexts. Parisse et al. (2017) propose the TEI format for oral and multimodal language corpora, pointing to ISO standard 24624:2016 as well as preconsiderations for the CMC-core schema, which has been proposed for corpora on computer-mediated communication (Luginbühl et al. 2021). Luginbühl et al. (2021: 2) specify CMC corpora interoperability for combined analysis on various CMC corpora, combining corpora of different types, and integrating CMC corpora into existing infrastructure, for example, as reasons to create and apply a TEI CMC schema.

These advantages (i.e., using a standard format and creating a sustainable workflow for higher corpus consistency) are also relevant concerning the FAIR principles (findable, accessible, interoperable, reusable; Wilkinson et al. 2016), which play an important role in research data management. While the principles of findability and accessibility depend mostly on the respective repositories in which the data are published, standardized metadata, which can be integrated in the TEI XML format in the designated metadata header, can help increase findability. Interoperability and reusability are improved by using standard data formats, such as TEI and ISO standard 24624:2016, and by higher consistency of the research data, which makes our proposed workflow a contribution to FAIRer corpus data.

3 MuMoCorp project: Additional requirements and lessons learned when realizing the corpus creation workflow

The workflow presented in this paper has been developed and tested on human-robot interaction data in the MuMoCorp project. The MuMoCorp project – Data Reuse of Multimodal and Multisensorial Corpora within the Dilthey Fellowship “Interaction & Space. From Conversation Analysis to Dynamic Interaction Models for Human-Robot Interaction” – prepares existing research data (see Pitsch 2016, 2020, 2023; Pitsch et al. 2016; Gehle et al. 2017) for long-term storage and further use as partly open data within the framework of an institutional repository. The rich data material is particularly interesting with respect to human-robot interaction and the multi-dimensionality of the interaction, and it includes different data formats such as videos, XML-based transcriptions, and robot log files. MuMoCorp's challenge is to organize and curate a large amount of data that has been collected, transcribed, and annotated over a period of roughly 10 years. Seven studies have been conducted exploring specific interactional features and procedures, and they have been stored as seven distinct but related subcorpora.

When curating this data and preparing it for reuse by other researchers, we encountered the following additional tasks, which constitute a prerequisite for doing so:

- Collecting and renaming the different files into a new, coherent data structure following a predefined specification
- Anonymizing the audiovisual recordings (see Krause, Ferger and Pitsch 2023a, 2023b)
- Collecting and structuring existing pieces of information about the participants, study details, and recordings as systematic metadata

We identified additional requirements for the corpus creation while using the workflow in practice. These may be specific to the particular project but might also be relevant for other projects and data:

- We wanted to keep the established workflows for transcribing and annotating the data, such as annotation tools (e.g., ELAN) and transcription/annotation conventions, unchanged.
- We wanted a version control for the data which, ideally, would not create any additional workload for researchers. Using Git/GitLab presented as a suitable solution for this task.
- A range of inconsistencies in the data can only be corrected manually. They should be corrected in the source files using the established tools (here: ELAN).
- A range of inconsistencies in the data can be checked automatically. These checks should be automated in a way that check results and where corrections are accessible continuously.

The data in MuMoCorp – stemming from a project on multimodal interaction and socio-technical situations – shares a range of features with computer-mediated communication, such as including various media types and system log files that can be similar to social media log files. Yet, there are also essential differences: In the MuMoCorp settings, the museum guide robot aims to act as an autonomous technical co-participant. The communication between the robot and the participants is not text-centred or internet-based. It occurs in real time and involves text-to-speech output, automatic speech recognition, head and arm movements for the robot, and verbal utterances and other modalities for the human participants. Therefore, the concept of *computer mediatedness* requires rethinking to appropriately grasp such constellations (for a start, see Arminen, Licoppe, and Spagnolli 2016).

4 A workflow to create machine-readable corpora for multimodal and computer-mediated resources

The workflow we propose for creating human- and machine-readable corpora can be applied both when starting a new project beginning with transcription/annotation or when processing an existing dataset for archiving, publication, or further analysis. It serves as our solution to the questions posed in chapter 1.

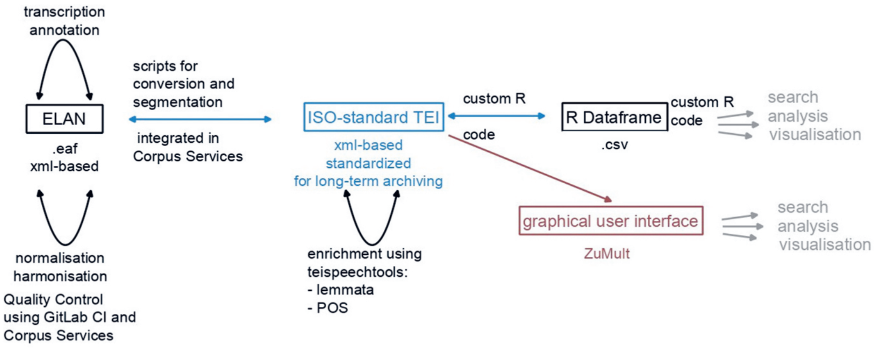


Figure 1: Workflow overview.

The workflow (see Figure 1) starts with (existing) transcription/annotation files which have been created with the ELAN editor (Sloetjes 2014; Max Planck Institute for Psycholinguistics 2020). These files could also be other XML-based source formats, such as EXMARaLDA files or TEI files that are manually generated or exported. To these source files we apply automated scripts for harmonization and normalization, as well as quality checks that require manual work for the harmonization (see chapter 4.2). From this source data, we export the data in the standard format ISO TEI. On the basis of the data available in the standard format, enrichments are added (e.g., lemmatization and part-of-speech tagging of verbal transcriptions), which can be continuously reapplied whenever changes are made to the data (see chapter 4.3). In a next step and as a basis for further analysis, the data – which has so far been normalized, harmonized, quality controlled, and enriched – can be automatically exported from the TEI files to various output formats (see chapter 4.4). The corpus data which is not available in text-based format (video files, etc.) are not altered by this workflow. These steps of harmonizing, checking, and converting the data are continuously reapplied using a dedicated system for version control (see chapter 4.1). This setup allows for dynamic analyses and visualizations to be performed continuously during the corpus creation process, which can help in the iterative development of annotation categories (see chapter 5), for example.

4.1 Automation using git, Continuous Integration, and GitLab

Answering our initial question (C) on how to ensure an automated, continuous control of the data consistency with our workflow, we used methods from software development for automated deployment that go beyond existing methods in continuous quality control (as in Hedeland and Ferger 2020; Ferger and Jettka 2021), namely Git version control and GitLab Continuous Integration (CI). Git version control[1] is the underlying version control system which allows for tracking and managing changes to data. GitLab[2] offers a platform and graphical user interface for the tracked changes and, with the CI functionality, allows automatically running scripts (e.g., validation techniques like schemas or specified consistency checks) on the tracked data without any user interaction. While git and GitLab are widely used, especially in research data management, other tools offer similar functionality, such as Apache Subversion or Gitea. Git versioning for text-based source data offers other benefits, such as the ability to revert to earlier stages of the files and the ability to track all changes to the files. Git for research data also makes research more reproducible and can be seen as best practice for research data management (see Hermann, Pietsch, and Cimiano 2021; Cyra, Politze, and Timm 2022; Erjavec, Kopp, and Meden 2023). Using the GitLab CI setup shown in Ferger, Krause, and Pitsch (2023), scripts are continuously applied when the data under version control is changed. This allows quality checks, correction, and data exports to be automatically reapplied whenever the source data is changed. The results of these checks are used for manually fixing these inconsistencies, but found inconsistencies do not prevent the corpus being used or changed further.

[1] https://git-scm.com/ (last accessed 14 February 2025).

[2] https://gitlab.com (last accessed 14 February 2025).

4.2 Data consistency

The initial question (B) concerns data consistency, or data quality as defined in Hedeland (2020), which calls for checking data for coherence and consistency. As can be seen in Figure 1, we aim to perform checks and automatic correction of inconsistencies on the XML-based source data, which in our case are ELAN files. An advantage of this approach is that further transcription, annotation, or manual harmonization of the files can be realized using the original tools and workflows. Since transcriptions and annotations are mostly carried out manually (and will continue to be for some time into the future, e.g., for multi-participant discussions and in dialects or less resourced languages), some inconsistencies can also only be resolved in a manual way.

To realize this approach, we used the Corpus Services Framework (Ferger et al. 2020; Hedeland and Ferger 2020) with additional extensions, including those for ELAN files (Arkhangelskiy, Hedeland, and Riaposov 2020), and adapted them to find and fix inconsistencies in the source files.[3] The generated list of identified inconsistencies including file names and locations of the inconsistencies can be used to facilitate manual correction of the files. Checks that have been applied include ELANTranscriptionChecker, which checks the adherence of verbal transcriptions to standardized conventions (here: GAT 2 [Selting et al. 2009]), ELANValidatorChecker, which checks if the ELAN XML file is valid according to ELAN specifications, ELANFileReferenceChecker, which checks if the linked media files exist, and ELANAnnotationChecker, which checks whether the annotation tiers adhere to annotation conventions – which is especially important for our coded interactional annotations.

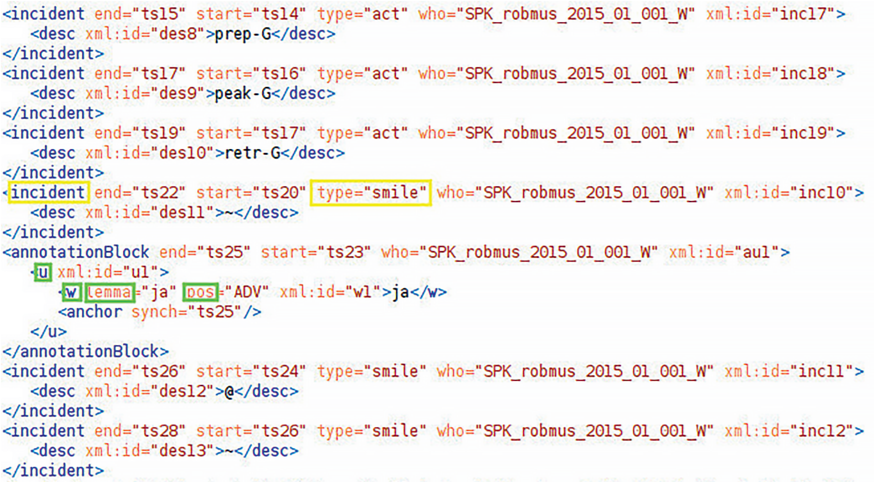
4.3 Modelling multimodal interaction using TEI as a standard and base format

The long-existing TEI guidelines, developed by the Text Encoding Initiative (TEI Consortium 2023),[4] are a standard for various text-based research fields. They are also adapted to spoken language, for example, with ISO standard 24624:2016 (hereafter referred to as ISO/TEI standard) "Language resource management – Transcription of spoken language" (Schmidt 2011; Hedeland and Schmidt 2022), which

[3] Our adapted version of the Corpus Services Framework is available at https://git.uni-due.de/mumocorp-open-access/corpus-services/ (last accessed 14 February 2025).

[4] For the history of the TEI see https://tei-c.org/about/history/ (last accessed 14 February 2025).

we see as an answer to our initial question (A). Similarly, the CMC-core schema for TEI (Beißwenger and Lüngen 2020) streamlines efforts in the CMC community to represent and structure corpora in a standardized way. While the TEI format is not yet widely used in conversation analysis and multimodal interaction research, there are exceptions such as Liégeois et al. (2015) and Parisse et al. (2017). One obstacle has been the amount of manual work required to generate it from multimodal interaction resources. To address this challenge in our proposed semi-automatic corpus creation workflow, we export the source data into TEI format and make the conversion accessible and reusable by including our export script in the Corpus Services Framework (see above). The script is based on exports adapted from the EXMARaLDA software suite (Schmidt and Wörner 2014). The TEI export includes a segmentation following the respective transcription conventions (here: GAT 2 [Selting et al. 2009]).



```
<incident end="ts15" start="ts14" type="act" who="SPK_robmus_2015_01_001_W" xml:id="inc17">
   <desc xml:id="des8">prep-G</desc>
</incident>
<incident end="ts17" start="ts16" type="act" who="SPK_robmus_2015_01_001_W" xml:id="inc18">
   <desc xml:id="des9">peak-G</desc>
</incident>
<incident end="ts19" start="ts17" type="act" who="SPK_robmus_2015_01_001_W" xml:id="inc19">
   <desc xml:id="des10">retr-G</desc>
</incident>
<incident end="ts22" start="ts20" type="smile" who="SPK_robmus_2015_01_001_W" xml:id="inc10">
   <desc xml:id="des11">~</desc>
</incident>
<annotationBlock end="ts25" start="ts23" who="SPK_robmus_2015_01_001_W" xml:id="au1">
   <u xml:id="u1">
      <w lemma="ja" pos="ADV" xml:id="w1">ja</w>
      <anchor synch="ts25"/>
   </u>
</annotationBlock>
<incident end="ts26" start="ts24" type="smile" who="SPK_robmus_2015_01_001_W" xml:id="inc11">
   <desc xml:id="des12">@</desc>
</incident>
<incident end="ts28" start="ts26" type="smile" who="SPK_robmus_2015_01_001_W" xml:id="inc12">
   <desc xml:id="des13">~</desc>
</incident>
```

Figure 2: TEI modelling example.

To model the data in our MuMoCorp project, we drew from modelling of spoken and computer-mediated resources in TEI format. We focused on the ISO/TEI standard for several reasons. This standard is recommended for long-term archiving by the Archive for Spoken German.[5] It is also used as an import format for the ZuMult corpus infrastructure (Frick and Schmidt 2020; Fandrych et al. 2022), which offers

[5] https://agd.ids-mannheim.de/uebernahme.shtml (last accessed 14 February 2025).

a graphical user interface to access corpus data, as well as in other tools such as WebLicht and WebMAUS (see Schmidt, Hedeland, and Frick 2021). For the verbal utterances of human participants in our data, no adjustments were needed. However, to use this standard for modelling multimodal aspects of the existing corpus data from human-robot interaction including novel data types, we needed to make some adaptations. These were carried out with the overarching idea of respecting the TEI standard such as to remain compatible with existing models, tools, and workflows. In particular, we used the following data model:

(a) Human utterances are represented by the <u> element, consisting of <w> for words with <pos> and <lemma> attributes, following the existing standard.

(b) To include multimodal annotations, such as bodily conduct and facial expressions, we used and adapted the TEI element <incident> (see Figure 2). According to TEI guidelines, this element "marks any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication" (TEI Consortium 2023). "Incident" is typically used to transcribe audible laughter in verbal transcription. We use the "type" attribute to refer to the level of (multimodal) annotation, such as "smile" or "nod", enabling analysis of the same phenomenon in different settings.

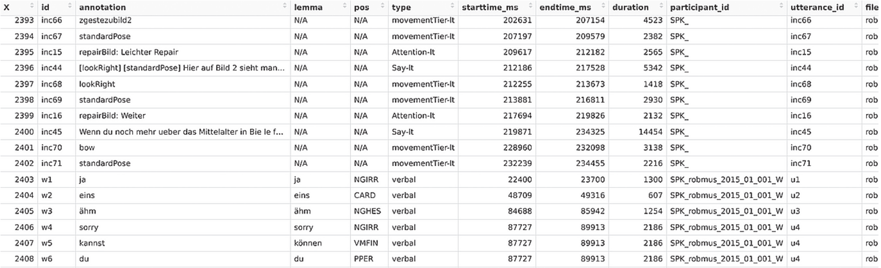
(c) The verbal utterances of the robot, which were generated via text-to-speech, were modelled similarly to multimodal annotations and not human verbal utterances because they do not share all the characteristics of human spoken language, such as interjections and prosodic aspects.

In comparison, the CMC-core defines its four basic units as spoken utterances, bodily activity, onscreen activities, and written utterances (Beißwenger and Lüngen 2020). Spoken or multimodal utterances in CMC-core are also represented using the <u> element; bodily activity is modelled using <kinesic>, and onscreen activities are modelled with <incident>. To improve the modelling of multimodal resources and its compatibility across settings and disciplines, it may be worth discussing if there was a benefit to instead use <kinesic> for bodily conduct and facial expressions. Here, further reflection, also in the light of other corpus data, might be helpful for future work.

We applied the Stuttgart-Tübingen Tagset (STTS) extension trained on spoken German on the FOLK corpus (Westpfahl and Schmidt 2013; Westpfahl et al. 2017) to the generated ISO-standard TEI files for participants' verbal utterances (and not the robots' text-to-speech output, which are not modelled as verbal utterances), using the teispeechtools library (Fisseni and Schmidt 2020), which employs TreeTagger (Schmid 1995). For internet-based resources, there are guidelines and a tagset for POS tagging presented in Beißwenger et al. (2015), which are relevant for internet-based CMC resources but which could not be applied in our context, such as emoticons or hashtags (which are not present in our type of resources). Since the tagset is also based on the STTS, the way of applying it can be identical.

4.4 DataFrame to facilitate analysis

Using the ISO-standard TEI as a source format allows various exports into different output formats, which answers our initial question (D). One goal was to generate a simple output with reduced complexity but that still accurately contains the relevant information of the source files, to facilitate analysis and visualizations in R. DataFrames for analyses in R are common for interactional linguistics, as in the tools ACT (Ehmer 2021, 2023) and EXMARaLDAR (Schürmann 2021) or for visualizations and statistics (Rühlemann 2020; Rühlemann and Ptak 2023). The existing tools did not satisfy our need to include all multimodal and sensor-based information in the dataFrame and its generation from a TEI file, so we developed a custom R script for this purpose, generating a dataFrame with columns inspired by those approaches, as seen in Figure 3. To allow for other programming languages and use cases, this dataFrame is additionally written into a csv table file.



X	id	annotation	lemma	pos	type	starttime_ms	endtime_ms	duration	participant_id	utterance_id	file
2393	inc66	zgestezubild2	N/A	N/A	movementTier-lt	202631	207154	4523	SPK_	inc66	rob
2394	inc67	standardPose	N/A	N/A	movementTier-lt	207197	209579	2382	SPK_	inc67	rob
2395	inc15	repairBild: Leichter Repair	N/A	N/A	Attention-lt	209617	212182	2565	SPK_	inc15	rob
2396	inc44	[lookRight] [standardPose] Hier auf Bild 2 sieht man...	N/A	N/A	Say-lt	212186	217528	5342	SPK_	inc44	rob
2397	inc68	lookRight	N/A	N/A	movementTier-lt	212255	213673	1418	SPK_	inc68	rob
2398	inc69	standardPose	N/A	N/A	movementTier-lt	213881	216811	2930	SPK_	inc69	rob
2399	inc16	repairBild: Weiter	N/A	N/A	Attention-lt	217694	219826	2132	SPK_	inc16	rob
2400	inc45	Wenn du noch mehr ueber das Mittelalter in Bie le f...	N/A	N/A	Say-lt	219871	234325	14454	SPK_	inc45	rob
2401	inc70	bow	N/A	N/A	movementTier-lt	228960	232098	3138	SPK_	inc70	rob
2402	inc71	standardPose	N/A	N/A	movementTier-lt	232239	234455	2216	SPK_	inc71	rob
2403	w1	ja	ja	NGIRR	verbal	22400	23700	1300	SPK_robmus_2015_01_001_W	u1	rob
2404	w2	eins	eins	CARD	verbal	48709	49316	607	SPK_robmus_2015_01_001_W	u2	rob
2405	w3	ähm	ähm	NGHES	verbal	84688	85942	1254	SPK_robmus_2015_01_001_W	u3	rob
2406	w4	sorry	sorry	NGIRR	verbal	87727	89913	2186	SPK_robmus_2015_01_001_W	u4	rob
2407	w5	kannst	können	VMFIN	verbal	87727	89913	2186	SPK_robmus_2015_01_001_W	u4	rob
2408	w6	du	du	PPER	verbal	87727	89913	2186	SPK_robmus_2015_01_001_W	u4	rob

Figure 3: DataFrame example.

The content in this dataFrame is exported directly from the TEI files and not normalized or changed, since we wanted to keep the harmonization in the source files. This is also important for dealing with the “Killer-Kriterium” (Schütte 2007: 71), a criterium for tools on how they deal with transcriptions during analysis. Many analysis tools process transcriptions in a way that they cannot be exported back into their source format after changes are made to them during analysis, which would make the transcriptions static and not dynamic. Non-dynamic transcripts would mean that after the analysis step, the transcriptions cannot be changed further, and findings in the analysis cannot be integrated easily into the source data, which is something we do want to utilize in our workflow. The continuous creation of this dataFrame helps with keeping the transcripts dynamic by delivering a dataFrame for up-to-date transcripts. Keeping the IDs as exact locations of certain elements in the source files allows for going back to the ELAN editor, for example, and changing things or automatically changing things in the source files based on the dataFrame. As the information of all transcripts comprising the corpus is grouped together in the dataFrame, more complex queries and analysis are made possible. Annotations and transcription in different ELAN files relating to the same video can thus be queried for simultaneity. Information on different levels of information can also be correlated, for example robot log files and verbal utterances of participants.

5 Application to other (CMC) resources

Answering question (C) and in terms of FAIR research data management, we have made the workflow available for reuse.[6] While we have discussed essential differences between the MuMoCorp resources and those of computer-mediated communication, many features of our workflow can be adapted to the specifics of CMC corpus data. In particular, the export of a dataFrame from CMC-core TEI files could be used for easier and more extensive analysis, as will be shown below.

6 Analysis of created corpora

To answer question (E) from our initial questions, the structured corpus facilitates more complex analyses than traditional corpus analysis constricted to verbal utterances for example. In what follows, this will be illustrated by an example analysis inspired by and adapted from Rühlemann (2018) and Rühlemann and Ptak (2023). To utilize aspects of human-robot interaction and sensor data as well as conversational analytic exploration, the example queries for actions of the robot and human reactions. An example for this is analyzing a movement of the robot along with non-verbal reactions of the participant in a study (e.g., the robot pointing to something in the room, and the participant nodding as a reaction; for detailed analysis

[6] The workflow and other resources are available at https://git.uni-due.de/mumocorp-open-access/ (last accessed 14 February 2025).

see Pitsch et al. 2016; Pitsch 2023). The first step is to formalize this phenomenon to allow for automatically finding all instances of it in the corpus. The phenomenon of robot movement could be formalized by querying the internal log file instructions of the robot, where movement is annotated as [pointLeftUp] or [pointRightUp]. The reaction of the participant can be formalized by manual annotations of the bodily conduct and facial expressions, such as “nod” or “smile”, or transcribed verbal utterances. So, one starting point of the analysis is an overview of all nodding annotations following in a temporal sequence of robot movement annotations. From there on, these instances can be classified using automated preliminary codes, which can be used for further formalizing (such as counting positive or negative reactions) or iterating manual annotation in order to find other instances of the phenomenon that were not technically identified. This iteration of manual or automatic annotation and analysis leads to dynamic transcriptions, since they are not frozen in one state after the analysis.

7 Conclusion

We have addressed the initial questions of this paper by presenting our reusable corpus creation workflow. By harmonizing inconsistencies directly in the source files and modelling data according to a TEI schema or standard, FAIRer and more sustainable research data can be created without too much additional work. Exporting a dataFrame to serve as the basis for complex analyses and visualizations facilitates the reuse of these analyses and visualizations as well. Using Git and GitLab CI capabilities not only allows for the measures for data quality to be applied continuously but also facilitates their reuse across resources. As any workflow is only as good as its output, we have shown that the machine-readable structured corpora generated can be queried by formalizing conversational analytic phenomena. Another benefit of using common standard formats is that the output of the workflow can be used for traditional corpus queries in corpus linguistics (Schmidt 2016) and as resources for second language teaching (Fandrych 2022). The advantages specified by Luginbühl et al. (2021: 2) for the TEI-based CMC core schema (interoperability for combined analysis on various corpora, combination of corpora of different types, and integrating corpora into existing infrastructure) can finally be seen as advantages of our workflow resulting in standardized outputs as well.

Acknowledgements

This project was financed by the Volkswagenstiftung (grant number 90886, PI: Karola Pitsch) and the Ministry of Culture and Science of the State of North Rhine-Westphalia (grant number PB22-076D).

References

Arkhangelskiy, Timofey, Hanna Hedeland & Aleksandr Riaposov. 2020. Evaluating and assuring research data quality for audiovisual annotated language data. In *Proceedings of CLARIN Annual Conference 2020,* Online Edition, 131–135.

Arminen, Ilkka, Christian Licoppe & Anna Spagnolli. 2016. Respecifying mediated interaction. *Research on Language and Social Interaction* 49 (4). 290–309.

Beißwenger, Michael, Thomas Bartz, Angelika Storrer & Swantje Westpfahl. 2015. Tagset und Richtlinie für das PoS- Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. *Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication* (EmpiriST 2015).

Beißwenger, Michael & Harald Lüngen. 2020. CMC-core: A schema for the representation of CMC corpora in TEI. *Corpus*. *Bases, corpus et langage* – UMR 6039 (20). https://doi.org/10.4000/corpus.4553.

Cyra, Magdalene Alice, Marius Politze & Henning Timm. 2022. A push for better RDM: Erfahrungsbericht aus dem Einsatz von git für Forschungsdaten. *Bausteine Forschungsdatenmanagement* 2. 1–17.

Ehmer, Oliver. 2021. act: Aligned Corpus Toolkit. R package version 1.2.2. https://cran.r-project.org/package=act (last accessed 14 February 2025).

Ehmer, Oliver. 2023. Arbeiten mit zeitalignierten multimodalen Korpora in R. Vorstellung des Aligned Corpus Toolkit (act). *Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion 24*. 67–126.

Erjavec, Tomaž, Matyáš Kopp & Katja Meden. 2023. TEI and Git in ParlaMint: Collaborative development of language resources. In *Selected papers from the CLARIN Annual Conference 2022*, 44–56.

Fandrych, Christian, Elena Frick, Julia Kaiser, Cordula Meißner, Annette Portmann, Thomas Schmidt, Matthias Schwendemann, Franziska Wallner & Kai Wörner. 2022. ZuMult: Neue Zugangswege zu Korpora gesprochener Sprache. In Heidrun Kämper & Albrecht Plewnia (eds.), *Perspektiven und Zugänge*, 305–312. Berlin & Boston: De Gruyter. https://doi.org/doi:10.1515/9783110774306-018.

Ferger, Anne, Hanna Hedeland, Daniel Jettka & Tommi Pirinen. 2020. Corpus Services. [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4725655.

Ferger, Anne & Daniel Jettka. 2021. Seamless integration of continuous quality control and research data management for indigenous language resources. In Monachini, Monica, Maria Eskevich (eds.), P*roceedings of CLARIN Annual Conference 2021*, Virtual Edition, 95–99.

Ferger, Anne, André Frank Krause & Karola Pitsch. 2023. A continous integration (CI) workflow for quality assurance checks for corpora of multimodal interaction. In Krister Lindén, Jyrki Niemi & Thalassia Kontino (eds.), *CLARIN Annual Conference Proceedings 2023*, 106–110. Leuven, Belgium.

Fisseni, Bernhard & Thomas Schmidt. 2020. CLARIN web services for TEI-annotated transcripts of spoken language. In Kiril Simov& Maria Eskevich (eds.), *Selected Papers from the CLARIN Annual Conference 2019*. Leipzig, 12–22.

Frick, Elena & Thomas Schmidt. 2020. Using full text indices for querying spoken language data. In Piotr Bański, Adrien Barbaresi, Simon Clematide, Marc Kupietz, Harald Lüngen & Ines Pisetta (eds.), *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora,* 40–46. Paris: European Language Resources Association. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-98143 (last accessed 14 February 2025).

Gehle, Raphaela, Karola Pitsch, Timo Dankert & Sebastian Wrede. 2017. How to open an interaction between robot and museum visitor? Strategies to establish a focused encounter in HRI. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (HRI '17), 187–195. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2909824.3020219.

Hedeland, Hanna. 2020. Towards comprehensive definitions of data quality for audiovisual annotated language resources. In Costanza Navarretta & Maria Eskevich (eds.), *Selected Papers from the CLARIN Annual Conference 2020*, 93–103.

Hedeland, Hanna & Anne Ferger. 2020. Towards continuous quality control for spoken language corpora. *International Journal of Digital Curation* 15 (1). 1–13.

Hedeland, Hanna & Thomas Schmidt. 2022. The TEI-based ISO standard 'Transcription of Spoken Language' as an exchange format within CLARIN and beyond. In *CLARIN Annual Conference*, 34–45.

Hermann, Fabian, Christian Pietsch & Philipp Cimiano. 2021. Conquaire infrastructure for continuous quality control. *Studies in Analytical Reproducibility: The Conquaire Project*. https://pub.uni-bielefeld.de/record/2951757 (last accessed 7 June 2021).

Hirschmann, Hagen & Thomas Schmidt. 2022. Gesprochene Lernerkorpora: Methodisch-technische Aspekte der Erhebung, Erschließung und Nutzung. *Zeitschrift für germanistische Linguistik* 50 (1). 36–81. https://doi.org/doi:10.1515/zgl-2022-2048.

Kendrick, Kobin H & Judith Holler. 2017. Gaze direction signals response preference in conversation. *Research on Language and Social Interaction* 50 (1). 12–32.

Kraft, Sophie, Angela Schmalen, Hendrik Seitz-Moskaliuk, York Sure-Vetter, Jennifer Knebes, Eva Lübke & Elena Wössner. 2021. Nationale Forschungsdateninfrastruktur (NFDI) e. V.: Aufbau und Ziele. *Bausteine Forschungsdatenmanagement* 2. 1–9.

Krause, André Frank, Anne Ferger & Karola Pitsch. 2023a. Detecting and tracking persons in video recordings of authentic social interaction: Analysis and anonymization. https://www.liri.uzh.ch/dam/jcr:f104d12e-416e-4246-8bca-dd16bd9808aa/CAMVA_2023_paper_1632.docx (last accessed 14 February 2025).

Krause, André Frank, Anne Ferger & Karola Pitsch. 2023b. Automatic anonymization of human faces in images of authentic social interaction: A web application. In Krister Lindén, Jyrki Niemi & Thalassia Kontino (eds.), *CLARIN Annual Conference Proceedings 2023*, 90–94.

Liégeois, Loïc, Carole Etienne, Christophe Benzitoun, Christophe Parisse & Christian Chanard. 2015. Using the TEI as a pivot format for oral and multimodal language corpora. *Text Encoding Initiative Conference and Member's meeting* 2015, Lyon, France.

Luginbühl, Martin, Vera Mundwiler, Judith Kreuz, Daniel Müller-Feldmeth & Stefan Hauser. 2021. Quantitative and qualitative approaches in conversation analysis: Methodological reflections on a study of argumentative group discussions. *Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion* 22. 179–236.

Max Planck Institute for Psycholinguistics, Nijmegen, The Language Archive. 2020. ELAN (Version 6.4) [Computer software]. https://archive.mpi.nl/tla/elan (last accessed 14 February 2025)

Mundwiler, Vera, Judith Kreuz, Daniel Müller-Feldmeth, Martin Luginbühl & Stefan Hauser. 2019. Quantitative und qualitative Zugänge in der Gesprächsforschung. Methodologische Betrachtungen am Beispiel einer Studie zu argumentativen Gruppendiskussionen. *Gesprächsforschung–Online-Zeitschrift zur verbalen Interaktion* 20. 323–383.

Parisse, Christophe, Céline Poudat, Ciara R Wigham, Michel Jacobson & Loïc Liégeois. 2017. CORLI: A linguistic consortium for corpus, language, and interaction. In Maciej Piasecki (ed.), *Selected papers from the CLARIN Annual Conference 2017*, Budapest, 15–24.

Pitsch, Karola. 2016. Limits and opportunities for mathematizing communicational conduct for social robotics in the real world? Toward enabling a robot to make use of the human's competences. *AI & SOCIETY* 31 (4). 587–593. https://doi.org/10.1007/s00146-015-0629-0.

Pitsch, Karola. 2020. Answering a robot's questions: Participation dynamics of adult-child-groups in encounters with a museum guide robot. *Réseaux* 220-221 (2-3), 113-150, https://doi.org/10.3917/res.220.0113.

Pitsch, Karola. 2023. Mensch-Roboter-Interaktion als Forschungsinstrument der Interaktionalen Linguistik. In Matthias Meiler & Martin Siefkes (eds.), *Linguistische Methodenreflexion im Aufbruch*, 119–152. Berlin & Boston: De Gruyter. https://doi.org/10.1515/9783111043616-005.

Pitsch, Karola, Timo Dankert, Raphaela Gehle & Sebastian Wrede. 2016. Referential practices. Effects of a museum guide robot suggesting a deictic 'repair' action to visitors attempting to orient to an exhibit. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 225–231. https://doi.org/10.1109/ROMAN.2016.7745135.

Pitsch, Karola, Anna-Lisa Vollmer, Katharina J. Rohlfing, Jannik Fritsch & Britta Wrede. 2014. Tutoring in adult-child interaction: On the loop of the tutor's action modification and the recipient's gaze. *Interaction studies. Social behaviour and communication in Biological and artificial systems* 15 (1). 55–98. https://doi.org/10.1075/is.15.1.03pit.

Rühlemann, Christoph. 2017. Integrating corpus-linguistic and conversation-analytic transcription in XML: The case of backchannels and overlap in storytelling interaction. *Corpus Pragmatics* 1 (3). 201–232. https://doi.org/10.1007/s41701-017-0018-7.

Rühlemann, Christoph. 2018. *Corpus Linguistics for Pragmatics: A guide for research* (1st edition). London & New York: Routledge. https://doi.org/10.4324/9780429451072.

Rühlemann, Christoph. 2020. *Visual linguistics with R: A practical introduction to quantitative Interactional Linguistics*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.228.

Rühlemann, Christoph & Matt Gee. 2017. Conversation analysis and the XML method. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 18. 274–296.

Rühlemann, Christoph & Alexander Ptak. 2023. Reaching beneath the tip of the iceberg: A guide to the Freiburg Multimodal Interaction Corpus. *Open Linguistics* 9 (1). 20220245. https://doi.org/doi:10.1515/opli-2022-0245.

Schmid, Helmut. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf (last accessed 28 March 2024).

Schmidt, Thomas. 2011. A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative*. Text Encoding Initiative Consortium (1). https://doi.org/10.4000/jtei.142.

Schmidt, Thomas. 2016. Construction and dissemination of a corpus of spoken interaction–tools and workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics* 31 (1). 105–132.

Schmidt, Thomas. 2018. Gesprächskorpora. In Thomas Schmidt & Marc Kupietz (eds.), *Korpuslinguistik*, 209–230. Berlin & Boston: De Gruyter. https://www.jstor.org/stable/j.ctvbj7k7n.13 (last accessed 6 April 2023).

Schmidt, Thomas. 2023. FOLK – Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch. *Korpora Deutsch als Fremdsprache* 3(1), 166–169. https://doi.org/10.48694/KORDAF.3737.

Schmidt, Thomas, Hanna Hedeland & Elena Frick. 2021. Ein Standard in der Praxis: ISO 24624:2016. Transcription of spoken language. FORGE 2021: Forschungsdaten in den Geisteswissenschaften – Mapping the Landscape – Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen (FORGE2021), Cologne. https://doi.org/10.5281/zenodo.5379639.

Schmidt, Thomas & Wilfried Schütte. 2010. FOLKER: An annotation tool for efficient transcription of natural, multi-party interaction. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). http://www.exmaralda.org/files/LREC_Folker.pdf (last accessed 14 February 2025).

Schmidt, Thomas & Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *Handbook on corpus phonology*, 402–419. Oxford. Oxford University Press. http://ukcatalogue.oup.com/product/9780199571932.do (last accessed 14 February 2025).

Schürmann, Timo. 2021. ExmaraldaR. https://github.com/TimoSchuer/ExmaraldaR (last accessed 14 February 2025).

Schütte, Wilfried. 2007. ATLAS.ti 5 – ein Werkzeug zur qualitativen Datenanalyse. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 8. 57–72.

Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg R Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock, Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10 353–402.

Sloetjes, Han. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.), *Handbook on corpus phonology*, 305–320. Oxford. Oxford University Press.

Stivers, Tanya. 2015. Coding social interaction: A heretical approach in conversation analysis? *Research on Language and Social Interaction* 48 (1). 1–19. https://doi.org/10.1080/08351813.2015.993837.

TEI Consortium (ed.). 2023. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. P5 Version 4.7.0. Last updated on 16th November 2023.* http://www.tei-c.org/Guidelines/P5/ (last accessed 14 February 2025).

Westpfahl, Swantje & Thomas Schmidt. 2013. POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. *Journal for Language Technology and Computational Linguistics* 28 (1), 139–153.

Westpfahl, Swantje, Thomas Schmidt, Jasmin Jonietz & Anton Borlinghaus. 2017. *STTS 2.0. Guidelines für die Annotation von POS -Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS)*. Working Paper. Version 1.1, März 2017. https://nbn-resolving.org/urn:nbn:de:bsz:mh39-60634 (last accessed 14 February 2025).

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (1). 160018. https://doi.org/10.1038/sdata.2016.18.