

Putting things on top of other things: The ZuMult platform for multimodal corpora and its ecosystem

Thomas Schmidt

University of
Duisburg-Essen

thomas@linguisticbits.de

Anne Ferger

University of
Duisburg-Essen

anne.ferger@uni-due.de

Elena Frick

Leibniz-Institute for the
German Language

frick@ids-mannheim.de

Abstract

We present ZuMult, an open-source platform for multimodal corpora, and its ecosystem. ZuMult has a flexible three-layer architecture that enables integration with different types of backend, and development of frontend applications tailored to the needs of highly diverse user groups. The paper puts a special emphasis on the way that ZuMult is built upon existing best practices, standards and technologies arguing that such an “ecosystem” makes it easier to integrate the technology with different workflows and technical environments.

1 Introduction

ZuMult (*Zugänge zu multimodalen Korpora gesprochener Sprache* ‘Access to multimodal corpora of spoken language’, Fandrych et al., 2023a, <https://zumult.org>) is an open-source¹ corpus platform for multimodal corpora. By ‘multimodal corpora’, we understand collections containing (at least) audio and/or video recordings of spoken language and corresponding transcriptions and fulfilling the “data maturity” conditions as suggested in Wamprechtshammer et al., (2022).²

As surveyed in more detail in Batinić et al. (2021) and Frick & Schmidt (2025), several multimodal corpus platforms have, in the last fifteen years, either been developed as new applications (e.g. DGD, Schmidt 2014, Spokes, Pezik 2015, LiRI Corpus platform, Grän et al. 2023, LAPIS, Pluschkovits et al. 2026) or as an extension or adaptation of platforms originally developed for written language (e.g. ANNIS, Krause & Zeldes 2016, CQPWeb, Love et al. 2017, KonText, Kren 2020), and a growing amount and diversity of audiovisual corpora are becoming available as FAIR compliant language resources in CLARIN (cf., for example, the CLARIN resource families “Spoken corpora”, “Multimodal corpora”, “L2 learner corpora”, “Oral History corpora”).

The most important challenge (and opportunity) for this kind of research data lies in its diversity, with respect to both characteristics of the data themselves (such as types of primary data, granularity and depth of annotation), and the multitude of ways in which they are used in empirical and applied research (ranging from interactional linguistics over dialectology to oral history studies or qualitative social sciences). The user study reported in Fandrych et al. (2016) has revealed that any sustainable technical solution in this area will therefore require considerable flexibility in backend (data) and frontend (applications) if it aims to be adopted more widely.

The ZuMult platform was designed with exactly this flexibility as a guiding principle, putting a special emphasis on support for approaches which are not predominantly quantitative in nature, but rely on fine-grained qualitative analysis of primary and secondary data.

ZuMult was developed in a project funded by the German Science Foundation between 2018 and 2022 as a flexible, versatile three-layer architecture allowing different types of data sources (e.g. file system, relational database) in the backend and different types of user-group specific adaptations (e.g. for language teaching, for conversation analysis, for variational linguistics) in the frontend.

In ZuMult’s project phase, the focus lay not only on the development of the architecture itself, but also on the development of frontend applications built upon it to address the needs of language teachers and learners (see the short descriptions of ZuMal, ZuViel and ZuRecht in Section 2). Currently, public instances of ZuMult exist at the Archive for Spoken German (AGD,

¹ The source code is available from <https://github.com/zumult-org/zumultapi> under a [GPL v3](https://www.gnu.org/licenses/gpl-3.0.html) license.

² There seems to be no widely established terminology to denote this kind of corpus. “Oral corpus” or “Audiovisual corpus” are also in use. In the CLARIN resource family “taxonomy” (<https://www.clarin.eu/resource-families>), we find “Spoken Corpora” and “Multimodal Corpora” as the tightest fits, but “Oral History Corpora” and “Disordered Speech Corpora” will usually also have audiovisual primary data. This may also be true of some “Parliamentary Corpora” and “L2 Learner Corpora”. See also Schmidt (2026) for an attempt at a classification of “non-textual language data”. More recently, ZuMult has also been applied for non-canonical written language data, more specifically: learner texts (Frick et al., 2025).

<https://zumult.ids-mannheim.de>), where the platform provides access to the largest corpora of spoken German (such as the Research and Teaching Corpus of Spoken German, FOLK, Schmidt, 2016, the corpus *Deutsch Heute* ‘German today’, documenting regional variation of spoken German, Brinckmann et al., 2008, and the corpus GeWiss, a comparable corpus of academic speech, Fandrych et al., 2012), and at the University of Texas in Austin (<https://tgdp-zumult.la.utexas.edu/>), where the data of the Texas German Dialect Project (TGDP) can be accessed via ZuMult (Boas et al., 2026, Schmidt et al. 2026). Further public instances are under construction for the French ESLO corpus (Enquêtes Sociolinguistiques à Orléans, archived at HumaNum, Schmidt, 2025), a number of multimodal and multisensorial corpora at the University of Duisburg-Essen (see Section 4) and for the EXMARaLDA demo corpus at the University of Hamburg.

ZuMult can be viewed as a piece of technology that both contributes to an infrastructure such as CLARIN, and profits from other technologies already integrated into or developed alongside the infrastructure. In the words of Monty Python (1970), ZuMult is thus a “thing that is put on top of other things”, and other things can be put on top of it (see Figure 9 in the conclusion).

The remainder of the paper will explain in more detail what this means: Section 2 briefly sketches the most important features of ZuMult from a user’s point of view. Section 3 then goes into more technical detail, emphasizing the interplay of different tools, standards, and services that underlies the architecture. Section 4, finally, gives an outlook on the current and future development of the platform.

2 Platform functionality

Generally speaking, a multimodal (spoken, audiovisual) corpus platform like ZuMult transfers functionality of written corpus platforms and extends it with functionality that is needed for and specific to the work with audiovisual language data and their transcriptions and annotations (Schmidt & Frick, 2025). Most importantly, this means that such platforms need to enable users:

- a) to browse through (read, listen to, view) the primary (audio/video) and secondary (transcribed/annotated) data,
- b) to select those (and only those) datasets that are relevant to their research questions,
- c) to visualise selected aspects of selected data for qualitative analysis, and
- d) to carry out, and possibly quantify, queries on transcriptions and their annotations with the option to re-contextualise query results (i.e. view them in the context of the underlying transcript and/or recording).

Metadata (on speech events and participating speakers) can guide or supplement all these processes (see Frick & Schmidt, 2025 for a comprehensive overview).

ZuMult’s architecture is designed in such a way that the required information (e.g.: What corpora are in the platform? What’s in corpus XY? Can I get transcript #123? What is the result of query Z?) can be delivered from one and the same backend to different frontends (“applications”), which are optimised for a specific usage scenario or a specific user group. The following subsections illustrate three such applications.

2.1 Data selection

The application ZuMal (*Zugang zu Merkmalsauswahl von Gesprächen*, ‘Access to the selection of conversational features’) supports the selection of suitable material for language teaching through an interface displaying selectors for relevant metadata and so-called “measures” reflecting the difficulty or degree of orality of a given dataset. Figure 1 illustrates how dialectal region (“Sprachregion”), vocabulary coverage (“Wortschatz”) and articulation speed (“Sprechgeschwindigkeit”) are used as selectors in the column on the left-hand side. Matching speech events are listed in tabular form and in a graphical visualisation on the right-hand side.

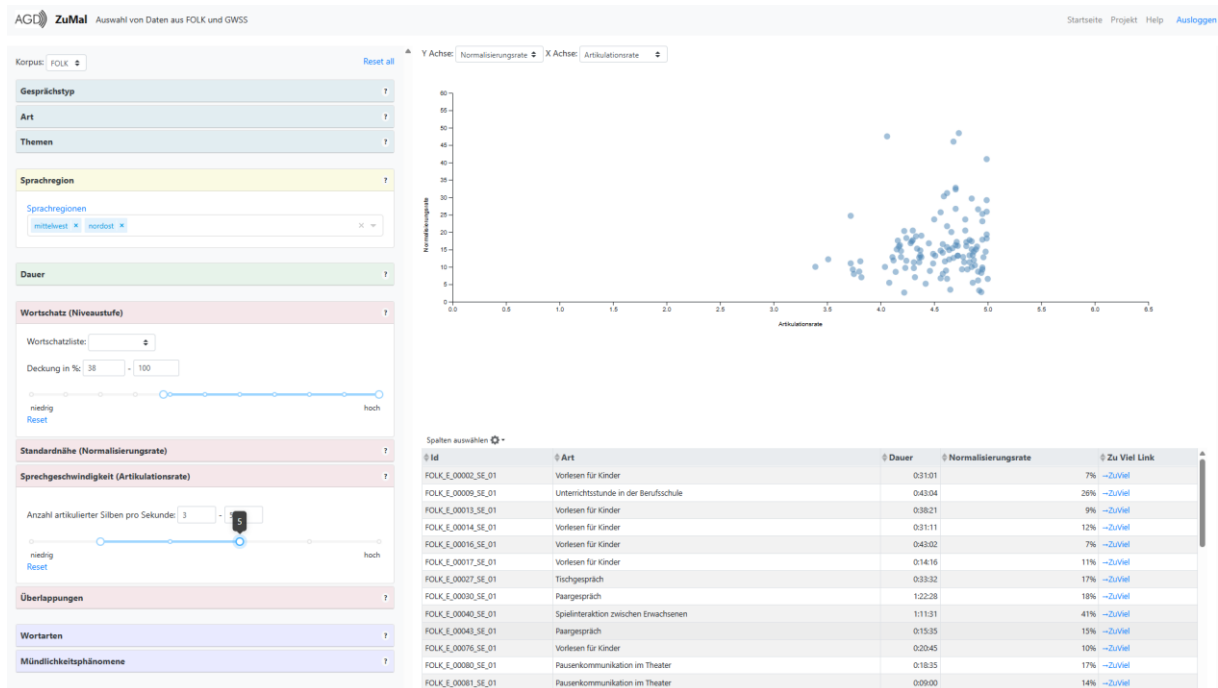


Figure 1: Data selection in ZuMal

Relevant metadata in that sense are, for example, the domain (private, institutional, public) and type (e.g. telephone conversation, workplace meeting, TV talk show) of interaction, and the (dialectal) region where it was recorded. Didactically relevant measures that can be automatically derived from the transcribed and annotated data are, among others, proximity to the standard (calculated as the ratio of tokens where the transcribed form deviates from standard orthography), articulation speed (syllables per second) or lexical coverage with respect to a reference wordlist (such as the Goethe wordlists corresponding to different levels of the Common European Framework of Reference for Languages CEFR).

The idea of the ZuMal tool is that language teachers can use it to find example material that is suitable to a given teaching situation, such as “Introducing B1 speakers of German to a private conversation in Bavarian that is sufficiently slow for them to understand”. Fandrych et al. (2023b) discuss such scenarios and ZuMal’s role in them in greater detail. Other means of data selection enabled by ZuMult include tabular overviews of speech event and speaker metadata that can be filtered and sorted, or the use of metadata inside search queries (see below).

2.2 Data visualization / Qualitative analysis

A crucial characteristic of many methodologies dealing with spoken and/or multimodal data is that, in contrast to classical corpus-linguistic approaches to canonical written data, they systematically include steps of close manual (i.e., intellectual) inspection of individual sections of primary and secondary data. Micro-sequential analysis, as practiced as a core method of conversation analysis (see, for instance, Sidnell, 2012), or the aforementioned use of excerpts in language teaching are prototypical cases in point. The application ZuViel (*Zugang zu Visualisierungs-Elementen für Transkripte* ‘Access to transcript visualisation features’, Schmidt et al., 2023, see Figure 2) is one among several tools developed in the ZuMult architecture to support this kind of approach.

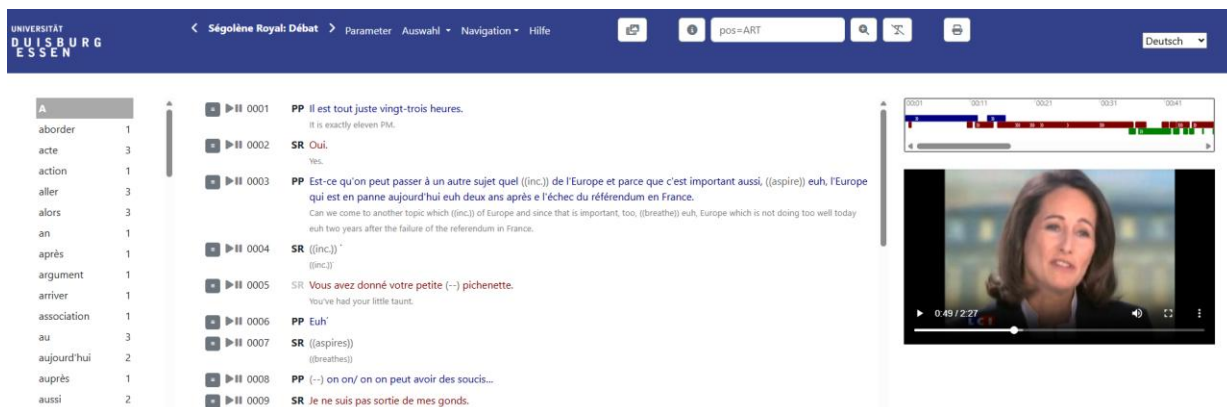


Figure 2: Interactive transcript display in ZuViel

ZuViel is an interactive transcript viewer which offers different, synchronised views (e.g. line transcripts, musical score transcript, lemma frequency list, density viewer, video with subtitles) of a transcript in which parameters (such as: display of transcribed vs. normalized forms, visualisation of speech rate) can be set by users according to their preferred perspective on the data. Interactive inspection and analysis of a piece of data is thus supported through functionality for foregrounding specific information (for example: highlighting all occurrences of tokens with a part-of-speech (POS) tag for pronoun) and for viewing relations between different types of data (e.g., synchronised media playback or a popup for speaker metadata). Whenever the functionality of the platform is not sufficient or appropriate for the analysis at hand, excerpts can be marked and downloaded in the formats for the most widely used desktop tools (such as Praat for phonetic analysis).

2.3 Query

With the application ZuRecht (*Zugang zu Recherchen auf Transkripten* ‘Access to transcript search’, Frick et al., 2023), the platform also provides an interface for querying audio and video transcriptions using the syntax of CQP (Corpus Query Processor Language, Evert et al., 2022). A notable aspect of this application is that it allows typical characteristics of spoken language to be included in searches. These can be alternative pronunciation variants, deviations from standardized forms, hesitations, pauses, and other non-verbal phenomena (e.g. laughter or blinking). Furthermore, time-aligned annotations, speaker overlaps and metadata can also be taken into account when querying interaction corpora. All these features can be easily integrated into ZuRecht CQP queries in form of token specifications or structural constraints, as illustrated by the query example in Figure 3. This CQP query will search in the German FOLK corpus for all pronunciation variants of ‘du kannst’ (you can), allowing for a possible hesitation (POS tag ‘NGHES’) or a pause in between. Both words should occur outside of speaker overlaps but within user action sequences (“as”) annotated with ‘Vorschlag/Angebot’ (suggestion/offer) and during vocational lessons (‘Unterrichtsstunde in der Berufsschule’).

```
([norm = "Du"] ([pos="NGHES"] | <pause/>)? [norm = "kannst"]
!within <speaker-overlap/>) within <as="Vorschlag/Angebot"/>
within <e se art="Unterrichtsstunde in der Berufsschule"/>
```

Figure 3: Complex CQP query

Thanks to the expressive power and flexibility of the CQP query language, the search via ZuRecht offers a wide range of possibilities, but it also requires some prior knowledge of the CQP syntax as well as the metadata and annotations available in the corpora. To make it easier for users to get started, a Query Builder has been integrated into the search input field of the AGD ZuMult instance. When the search field is clicked, the Query Builder automatically suggests syntactic elements that are valid at the current position in the query and allows users to select from the annotation and metadata categories as well as their possible specification values available in the selected corpus (Figure 4).

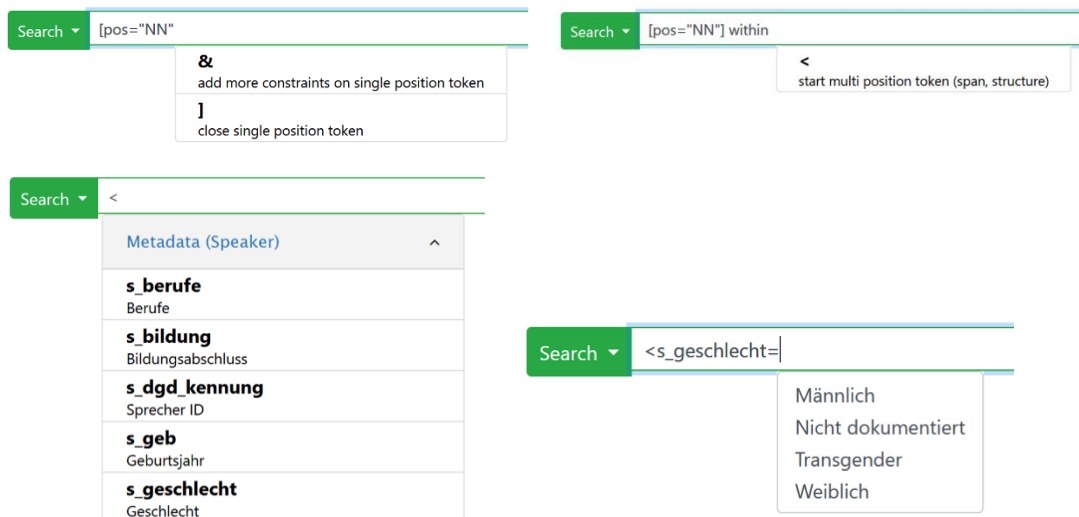


Figure 4: The query builder displaying help for CQP syntax (top left, top right), a suggestion box with the available categories of speaker metadata (bottom left) and a suggestion box with the available values for the metadata category ‘Gender’ (bottom right).

ZuRecht is primarily a CQP-based KWIC concordancer with special features (e.g. audio playback for KWIC lines, inclusion of video stills, see Schmidt & Frick 2025 and Figure 5) for audiovisual data. Going beyond that basic functionality, it was further extended to enable searches for repetitions (Frick et al., 2024) and n-grams that can be executed directly on the search index without the need of extra annotations or n-gram-lists. To this end, dedicated search forms have been implemented in ZuRecht that allow users not only to search for repetitions and n-grams, but also to formulate queries considering the specific properties of spoken language. For example, users can specify whether hesitations should be ignored or taken into account when computing n-grams (e.g. whether *good erm afternoon* should yield one or two bigrams). As another example, when searching for repetitions, users can define whether a speaker’s utterance is repeated by another speaker or if they repeat their own utterance, and whether such self-repetitions occur within the same turn or only after a speaker change – possibly triggered by an intervention from another speaker, such as a question or a request for clarification.

Insgesamt: 51						
1	EXB_Anne_Will	AW	... damit sich ratzfat	auf	Seite eins katapultiert ...	▶ ◀ 🔍 📄
2	EXB_Anne_Will	GL	... könnten doch mal	über	das Wahlrecht nachdenken ...	▶ ◀ 🔍 📄
3	EXB_Anne_Will	GL	... was die Bildzeitung	auf	Seite eins Gut ...	▶ ◀ 🔍 📄
4	EXB_Anne_Will	AW	... Sie wollten nur	auf	die Seite eins ...	▶ ◀ 🔍 📄
5	EXB_Anne_Will	GL	... bei der Bildzeitung	auf	Seite eins gekommen ...	▶ ◀ 🔍 📄
6	EXB_Anne_Will	GL	... die gesamte Zeit	über	Hartz vier Sätze ...	▶ ◀ 🔍 📄
7	EXB_Anne_Will	HG	... da natürlich nicht	auf	die Schulter klopfen ...	▶ ◀ 🔍 📄
 00:01:38.51 00:01:39.06 00:01:39.62 00:01:40.18 00:01:40.74 00:01:41.30						
8	EXB_Beckhams	VIC	... really good friends ,	on top of	everything . W do ...	▶ ◀ 🔍 📄
	0041	VIC	((0.3s)) You know he was sitting there with his family and/ and I really liked that.			
	0042	VIC	And he is a very kind ((0.5s)) person.			
	0043	VIC	And we/ we are really good friends, on top of everything.			
	0044	PAR	W/ do/ what's been the lowest?			
	0045	PAR	That's the/ the nice part of the marriage.			
9	EXB_HartAberFair	RP	... Dieselpreisentwicklung natürlich auch	auf	ah uns ausgewirkt ...	▶ ◀ 🔍 📄

Figure 5: A KWIC concordance in ZuRecht displaying video stills and the transcript excerpt corresponding to line 8.

3 ZuMult and its ecosystem: Things on top of other things

The starting point for ZuMult’s design and architecture was experience gained in the development of EXMARaLDA (Schmidt & Wörner, 2014) and the Database for Spoken German (Schmidt, 2014) as well as in the compilation of the corpora FOLK and GeWiss (see introduction). Those tools and corpora, in turn, have greatly profited from (and partly contributed to) standards and best practices developed in the last 30 years, both on a larger scale in text and media technology (Unicode, XML, the MPEG-family of standards), and in more specific linguistic contexts (e.g. Schmidt et al., 2009; Schmidt et al., 2010; Westpfahl & Schmidt, 2016). ZuMult reuses ideas and concepts that have proven fruitful in long-standing application of these resources and attempts to add and improve things that have turned out to be missing, deficient or hard to use.

3.1 Data model / Data interoperability

ZuMult is built on top of an abstract data model which relates corpora, events, speech events and speakers as conceptual objects with binary media objects (audio and video files) and textual transcription objects as sketched in Figure 6.

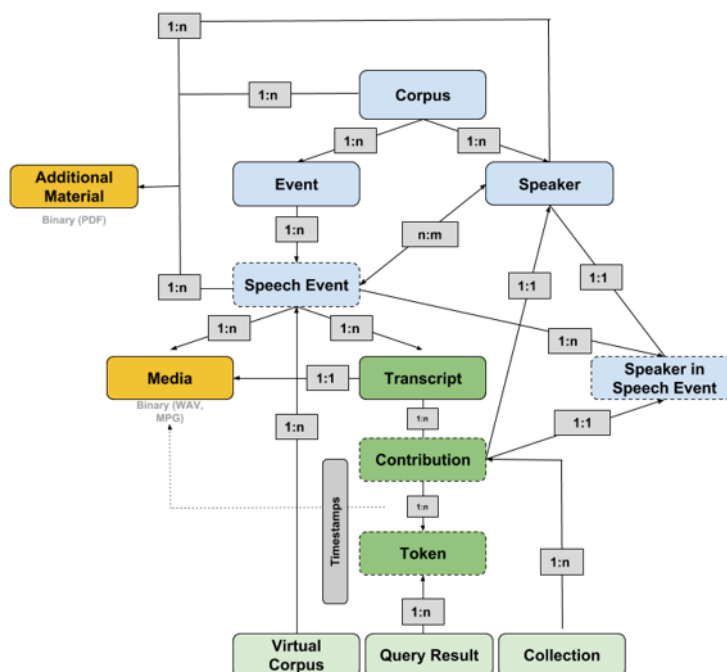


Figure 6: ZuMult data model for corpora

On a more concrete level, ZuMult reuses EXMARaLDA’s data model and API with its extensive interoperability, meaning that, from the outset, ZuMult is built on top of a technology interoperable with the most widely used tools (ELAN, Praat, EXMARaLDA, FOLKER, CLAN, Transcriber and others) and formats for multi-layer annotation of audiovisual data (Schmidt et al., 2009).

More recently, this interoperability has been extended to include output formats of ASR technology (e.g. the JSON format of OpenAI Whisper). On top of this level of exchangeability of tool formats sits a standardisation via “ISO 24624:2016 Language resource management – Transcription of spoken language” which itself rests on top of the long-standing efforts of the Text Encoding Initiative. This standard is currently the best option to represent both the multi-tier time-based “macro” structure of tools like ELAN or EXMARaLDA and the hierarchical token-based “micro” structure of annotated transcription text in a single document (Schmidt, 2011). It has been established as a standard in the CLARIN infrastructure (Schmidt et al., 2017; Fisseni & Schmidt, 2020; Hedeland & Schmidt, 2022), and its documentation and tool support are currently improved in a dedicated cooperation project (“Transcription+”) in the Text+ consortium of the German NFDI³.

Concerning metadata (on speech events, on speakers, on the corpus as a whole), no comparably stable or established basis for interoperability and standardisation exists. ZuMult’s data model in this

³ See <https://linguisticbits.de/transcription-plus.html> for a preliminary website

respect therefore remains on a relatively abstract level which only requires metadata to be organised as attribute-value pairs underneath a ‘corpus > event > speech event’ hierarchy and a speaker list with a n:m assignment between speakers and speech events (see Figure 6 and Schmidt, 2014b). On the implementation level, ZuMult so far has been made to work with the metadata schema of the Archive for Spoken German (Dickgießer, 2017) and with the metadata schema of the EXMARaLDA Corpus Manager (COMA, Wörner, 2012). If the need arises, it could also be made to work with IMDI, individual CMDI schemas or metadata encoded in TEI headers. We found, however, that these solutions are not frequently used in practice to encode the actual corpus metadata (the single most common solution for this being spreadsheet tables that we convert to the COMA format for integration into ZuMult). Rather, they are in more wide-spread use only for the purpose of documenting selected metadata for archiving purposes.

3.2 Workflows / Processing Pipelines

Beyond the level of formats, ZuMult is compatible with (and thus can be made to sit on top of) workflows developed for larger corpus projects. A ZuMult integration has been established as a final step of the workflows for the ongoing compilation of FOLK (Schmidt, 2016) and the TGDA (Boas et al., 2026), and, most importantly, is now also an element in the git-based continuous integration workflows described in Ferger et al. (2023). In the latter form, using the ZuMult platform also becomes an option in more dynamic setups, i.e. contexts in which corpora do not follow a strict release schedule, but are built up in smaller research projects in a fluid iteration of data collection, curation and analysis. This allows smaller and ongoing corpus projects to leverage ZuMult’s functionality.

In the git-based continuous integration workflows, several automated steps support the integration into ZuMult by checking, enriching, converting and indexing the elements of a corpus. In addition to the steps of harmonizing and repairing inconsistencies, which were already established in the workflows of the INEL project (Hedeland & Ferger, 2020), recently added workflow steps include:

- a conversion of transcription files (i.e. from ELAN or EXMARaLDA files, or results from Whisper ASR) into the standardised TEI format required by ZuMult,
- tokenisation and the addition of annotations on the token level, such as part-of-speech, lemmatization, orthographic normalisation, language detection,
- phonetic annotation (with the G2P web service, Reichel & Kisler 2014) and speech rate annotation (syllables per second, calculated from time-alignment and G2Ps syllabification),
- automated translation using the DeepL API (<https://developers.deepl.com/>),
- compilation of metadata from different sources, e.g. spreadsheets or automatic extraction of media metadata via FFPROBE (<https://ffmpeg.org/ffprobe.html>),

A final step of the workflows is the (Lucene) indexing of the transcription and metadata files for ZuMult. A schematic depiction of a workflow is shown in Figure 7.

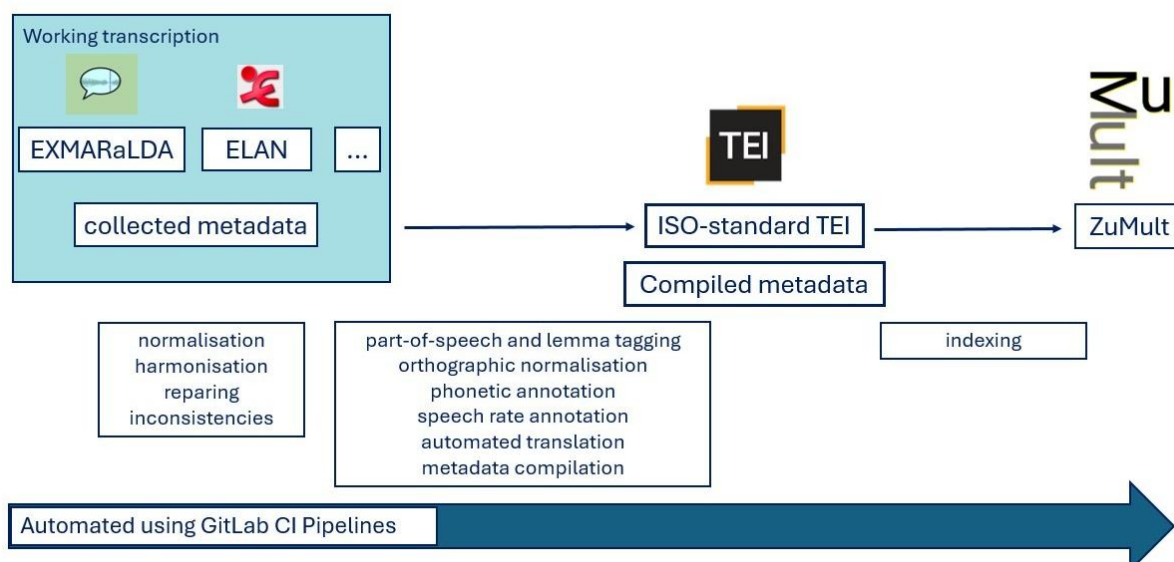


Figure 7: Continuous workflows for ZuMult integration.

Similarly to the integration of corpora into ZuMult, the build processes and deployment of the platform itself are automated using continuous integration workflows alongside Maven and Docker as distribution methods for the code, thus facilitating integration into different environments as well as improving software sustainability.

3.3 Query implementation

ZuMult’s query capabilities (see Figure 9 in the conclusion) are developed on top of the Multi-Tier Annotation Framework (MTAS, Brouwer et al., 2016) which itself is built on top of the industry level search software Lucene. MTAS can read and index transcription files adhering to ISO 24624:2016, and make them queryable via the CQP syntax, which has been identified (Fandrych et al., 2016) as the most promising candidate for a query language that is both widely known and suitable for the complexity of spoken language data.

MTAS is an open-source framework originally developed within the Nederlab project for working with large collections of digitized texts with multi-layer annotations. In the ZuMult project, this framework was applied for the first time to index and search audiovisual interaction corpora (cf. Frick & Schmidt, 2020). In what follows, several aspects of the integration of MTAS into the ZuMult architecture are described.

Adjusting the configuration file of the MTAS document parser: The configuration file of the MTAS document parser defines a set of mapping rules that determine which elements of ISO 24624:2016 transcripts should be stored in the search index and how they can be addressed later in CQP queries. These rules can specify, for example, whether annotated pauses and other non-verbal elements of spoken transcriptions should be ignored, treated as properties of words, or indexed as independent tokens. In addition, the mapping rules allow for the use of various filters and constraints to select only specific elements within a group or to modify them dynamically. This functionality can be used, for instance, to store word forms both in their original form and in lowercase, that would enable case-insensitive searches. Another example is the reduction of storage requirements in the search index by excluding speech rate annotations with the tag “0.0” labeling contributions that contain only non-verbal elements. Furthermore, integrating speaker and speech event metadata into transcripts (in the same way as other span-based annotations) and storing them in the search index together with corpus transcriptions and annotations enables metadata to be used as constraints directly in CQP queries (as demonstrated in the example in Section 2.3).

Indexing of transcript documents: Indexing of transcript documents is generally dependent on the specific ZuMult instance and varies according to the type of corpus data and the underlying user requirements. For example, all data may be stored in a single search index, or multiple complementary or alternative search indices may be created and used depending on the system goals. Because the corpora in the AGD collection vary considerably with respect to transcription conventions, the cross-corpus queries are rarely required, therefore a separate index file was created for each corpus. This approach keeps index sizes manageable and improves retrieval performance.

Corpora whose transcripts include punctuation marks are indexed in two different ways: once with punctuation and once without. This allows users to choose between two search modes and to specify token distances in their queries either with or without taking punctuation into account. In addition, two index types are created for each corpus: a transcript-based index, which indexes tokens according to their sequential position in the transcript document, and a speaker-based index, which processes tokens separately for each speaker while treating the tokenizations of other speakers as annotations of the current speaker. This approach helps to resolve conflicting tokenizations (cf. Sauer & Lüdeling, 2016) in spoken corpora that arise from speaker overlaps (for more details see Section 3.2 in Frick et al., 2022; Section 4.2. in Frick et al., 2023 and Section 5.4. in Frick & Schmidt, 2025).

Implementation of the search component: The search itself is performed via Java API for RESTful Web Services (JAX-RS), which establishes a connection to the ZuMult backend. Within the ZuMult backend, an instance of the *Searcher* class is created for each request. This class is responsible for preparing the search request before it is forwarded to the search engine and for post-processing the search results. For example, it determines which indices are to be queried and enriches the search results with various types of metadata, such as the time at which the search was executed.

The search engine then handles the parsing of the CQP query string, executes the query against the search index, and retrieves the required information from the index, such as the token IDs of the

matches or associated metadata that may later be needed for further sorting of the results. In addition, the search engine of the AGD ZuMult instance has been extended to support searches for repetitions and n-grams. For this purpose, the search engine performs several checks directly on the search index during query execution, such as whether both elements of a repetition originate from the same contribution or whether a speaker change occurs between the respective token positions.

Presentation of the search results: The ZuMult backend can return three types of search results. Each result type is modelled as an object within the ZuMult architecture and is implemented by a dedicated class with a specific interface. The *SearchResult* class contains only the number of hits and the number of transcripts in which they occur, along with selected metadata about the search process. *SearchResultPlus* additionally includes the token IDs of the matches and further metadata for each hit. The *KWIC* class is responsible for transforming search results into a KWIC object and for retrieving the relevant contextual information for each hit from the transcript files.

The subsequent serialization of the KWIC object into XML format is handled by the web services layer. Depending on the intended use of the search results – for example, whether they are requested by a ZuRecht instance (Section 2.3) or by the CLARIN Federated Content Search (FCS, Körner et al., 2024) – different XML serializations of the results may be produced.

Although MTAS is currently the search engine used in all active ZuMult instances, the ZuMult architecture is designed in such a way that the Search API can be extended at any time to include additional search engines. The use of multiple search engines and the flexible switching between them, should allow users to leverage the strengths of different query languages. Currently, it is being evaluated whether the BlackLab search engine (Does et al., 2017) is a suitable candidate to extend the ZuMult Query component. Like MTAS, BlackLab is capable of handling data in the ISO 24624:2016 format and can be adopted for new environments via a configuration file. Both MTAS and BlackLab are implemented in Java and built on the Apache Lucene information retrieval library, which have a large user and developer community as well as a strong long-term maintainability, providing a sustainable technical base for the ZuMult architecture.

3.4 Other components

Other components of ZuMult also rest on top of existing technology: For instance, for fine-grained multimodal analysis, transcript excerpts are force-aligned with their audio on the phoneme level using the CLARIN service WebMaus (Kisler et al., 2017) and still images extracted from the corresponding video file via the FFMPEG suite (<https://ffmpeg.org/>). This is supplemented with an F0 pitch contour extracted from the audio using Praat scripts and visualised as an SVG drawing (see Figure 8).

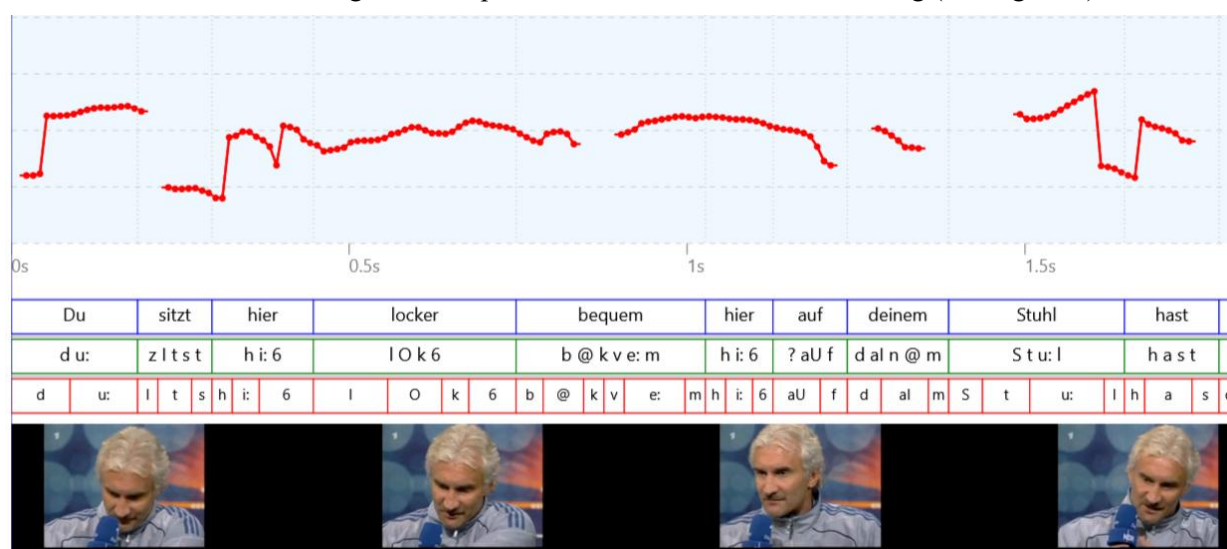


Figure 8: A micro-view of a transcript excerpt in ZuMult: phoneme-level alignment achieved via MAUS web service, pitch contour calculated via Praat.

3.5 Interoperability with CLARIN

Last but not least, data published on an instance of the ZuMult platform can be directly integrated into the CLARIN infrastructure. This is the case, for example, when the platform obtains its actual data (transcripts, audio, video, metadata) from a CLARIN repository, as is the case in the ongoing

development of a ZuMult instance for the ESLO corpus (Schmidt, 2025) which is tied to the Cocoon platform of the French HumaNum infrastructure (<https://cocoon.huma-num.fr/>).

Additionally, ZuMult is currently extended to serve CLARIN protocols for FCS and become an endpoint for an FCS aggregator. The Transcription+ project (see above) will demonstrate this by setting up an FCS endpoint for the EXMARaLDA demo corpus (<http://doi.org/10.25592/uhhfdm.8364>) in collaboration with the CLARIN/Text+ centres in Hamburg.

4 Current and future developments

Current development of ZuMult is carried out at the University of Duisburg-Essen in two contexts:

- The ZAT project (<https://zat.nrw/>) deals with assistive technologies. Corpus data in this context comprise studies of human-machine interaction (e.g. people interacting with a multimodal chatbot or a remote instruction robot), which are recorded, not only on video from a number of perspectives, but also with sensor technology (eye trackers, biometrical sensors) and through machine log files.
- The research group on personal reference (<https://forschungsgruppe-pronomen.de/en/>) studies the use of personal, indefinite, and demonstrative pronouns used in German to refer to present and absent persons. One corpus in this context consists of data collected at emergency medical drills (“mass casualty incidents”) in the form of videos from many perspectives, eye tracking data from main actors, and (perspectively) GPS data on actors’ positions in the scene.

Challenges for ZuMult in both contexts include the handling of very large amounts of video data, the integration and synchronized visualisation of multi-modal annotations and additional time-series data (often with spatial coordinates), and the dynamic (often opportunistic) ways these data are processed, reflecting an orientation towards immediate needs of data exploration and analysis in the research process, rather than towards systematic long-term corpus building. The tighter integration of ZuMult with CI pipelines and the measures aimed at an optimization of the distribution of the platform described in Section 3.2, as well as the tools for fine-grained multimodal analysis illustrated in Section 3.4 are largely guided by these requirements. Additionally, we are exploring ways of interfacing ZuMult with the R programming language to exploit this framework’s strengths in statistical analysis and data visualisation.

Prospectively, we are envisaging a second phase for the project to further develop the basic architecture of the platform in three different ways:

- an extension and generalization of the data model to (a) allow for more fine-grained processing (visualisation, navigation) of data, (b) cater also for non-canonical written data (e.g. learner texts) and (c) enable the integration of any kind of time-series data alongside audio, video and transcriptions.
- an extension of functionality for persistent collaborative work, enabling users to create, store and share individual collections of data inside the platform, and to enrich corpora with user-generated comments, references and annotations
- an exploration of LLM-based techniques for automatically enriching data and supporting queries

In the mid-term, we thus hope to extend ZuMult’s user base and further establish it as a technology in different research communities dealing with audiovisual language data.

5 Conclusion

As Figure 9 illustrates, ZuMult, as a piece of technology, is built on top of lots of other pieces of technology which, in turn, are also built on top of each other, and it can be integrated in and with other solutions. The fact that the spread of best practices and the establishment of standards for the relevant technology is advancing, and that ties between such solutions are being strengthened, makes an undertaking such as ZuMult easier to implement, easier to maintain and, hopefully, relevant to a larger community of users.

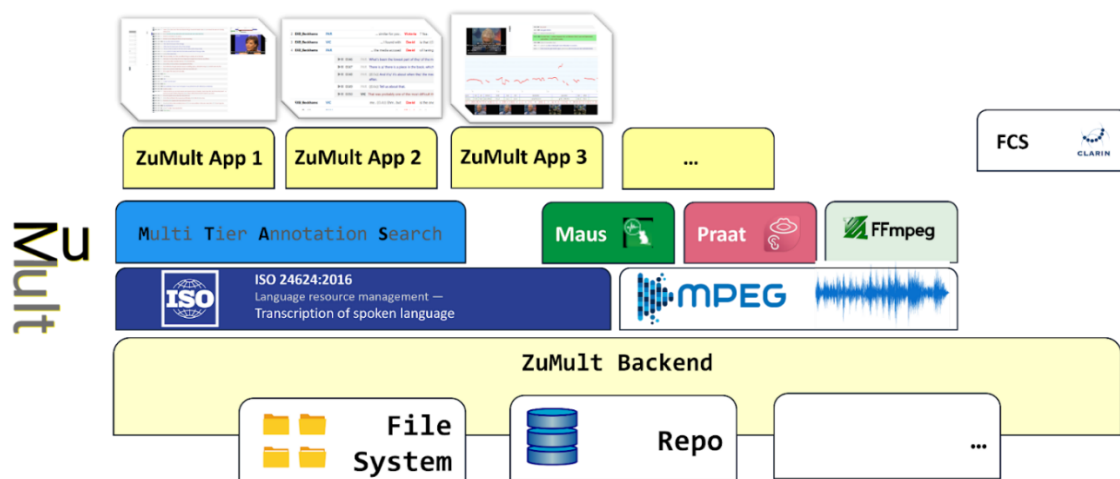


Figure 9: ZuMult and the things on top of which it is built

In our current work, we observe especially that, compared to large “monolithic” solutions, building a platform with such a modular approach makes it more manageable also for smaller projects to adapt and run locally. In the medium-term, this may have a positive effect on the community’s ability to contribute to an infrastructure like CLARIN.

Acknowledgments

The original ZuMult project was funded by the German Research Council (DFG) from 01/12/2018 to 31/03/2022 (project number 359880971) as a collaborative project of the Leibniz-Institute for the German Language (program area ‘Oral corpora’, Thomas Schmidt), the University of Leipzig (Herder-Institute, Christian Fandrych) and the University of Hamburg (Hamburg Centre for Language Corpora, Kai Wörner).

Current work at the University of Duisburg-Essen is carried out with funding from the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen for the project „Zentrum Assistive Technologien Rhein-Ruhr, ZAT Rhein-Ruhr“ (PB22-076D) and from the DFG for the project “Multimodalität von Personenreferenz: Pronominale Personenreferenz in Notfallübungen von Mediziner*innen und Feuerwehr” as part of the research group "Praktiken der Personenreferenz: Personal-, Indefinit- und Demonstrativpronomen im Gebrauch" (FOR 5317), project number 457855466. Karola Pitsch is the principal investigator on both projects.

The Transcription+ project is a cooperation project of the Text+ consortium, funded by the DFG (project number 460033370) as part of the German National Research Data Infrastructure (NFDI). Thomas Schmidt is the principal investigator. The Academy of Sciences and Humanities in Hamburg and the University of Hamburg are supporting the project in Text+ integration.

We would also like to gratefully acknowledge the ongoing collaborations with the University of Texas in Austin (Hans C. Boas, Margaret Blevins) and with the Laboratoire Ligérien de Linguistique at the University of Orléans (Flora Badin, Lotfi Abouda, Céline Dugua, Marie Skrovec).

References

- Batinić, J., Frick, E. & Schmidt, T. (2021): *Accessing spoken language corpora: an overview of current approaches*. In: Corpora Vol. 16, Issue 3. Edinburgh: Edinburgh University Press, 2021. 417-445. <https://doi.org/10.3366/cor.2021.0229>
- Boas, H., Schmidt, T. & Blevins, M. (2026, accepted). *TGDA 2.0 - A corpus platform for the Texas German Dialect Archive*. Language Resources & Evaluation. Springer.
- Brinckmann, C., Kleiner, S., Knöbl, R. & Berend, N. (2008). *German Today: an areally extensive corpus of spoken Standard German*. Proceedings of LREC 2008, Marrakech, Morocco. European Language Resources Association (ELRA), 2008. S. 3185-3191.
- Brouwer, M., Brugman, H. & Kemps-Snijders, M. (2016). *MTAS: A Solr/Lucene based Multi-Tier Annotation Search solution*. Selected papers from the CLARIN Annual Conference 2016. Aix-en-Provence, 19-37.
- Dickgießer, S. (2017). *Metadaten im Programmbereich „Mündliche Korpora“ des IDS*. Mannheim: Leibniz-Institut für Deutsche Sprache. <https://nbn-resolving.org/urn:nbn:de:bsz:mh39-62525>

- Does, J., Niestadt, J. & Depuydt, K. (2017). *Creating research environments with BlackLab*. Odijk, J. & van Hessen, A. (eds.) CLARIN in the Low Countries, 151-165. London: Ubiquity Press. DOI: <https://doi.org/10.5334/bbi>
- Evert, S. & CWB Development Team (2022). *The IMS Open Corpus Workbench (CWB) - CQP Interface and Query Language Manual CWB Version 3.5*. https://cwb.sourceforge.io/files/CQP_Manual.pdf
- Fandrych, C., Meißner & C., Slavcheva, A. (2012). *The GeWiss corpus: Comparing spoken academic German, English and Polish*. Schmidt, T. & Wörner, K. (eds.): Multilingual Corpora and Multilingual Corpus Analysis. Hamburg Studies in Multilingualism 14. Amsterdam: John Benjamins, 319-338.
- 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*. Proceedings of LREC 2016, Portorož, Slovenia. Paris: ELRA, 280-287.
- Fandrych, C., Schmidt, T., Wallner, F. & Wörner, K. (eds.) (2023a). *Zugänge zu multimodalen Korpora gesprochener Sprache*. Special issue of Korpora Deutsch als Fremdsprache (3) 1. <https://kordaf.tu-journals.ulb.tu-darmstadt.de/issue/92/info/>
- Fandrych, C., Meißner, C., Schwendemann, M. & Wallner, F. (2023b). *ZuMal: Zielgruppenspezifische Gesprächsauswahl aus Korpora gesprochener Sprache*. In: Fandrych et al. (2023a), 13-43.
- Ferger, A., Krause, A. & Pitsch, K. (2023). *A Continuous Integration (CI) Workflow for Quality Assurance Checks for Corpora of Multimodal Interaction*. Lindén, K. / Niemi, J. / Kontino, T. (eds.), CLARIN Annual Conference Proceedings 2023. Leuven, 106-110.
- Fisseni, B. & Schmidt, T. (2020). *CLARIN Web Services for TEI-annotated Transcripts of Spoken Language*. Simov, K. & Eskevich, M. (eds.): Selected Papers from the CLARIN Annual Conference 2019. (= Linköping Electronic Conference Proceedings 172). Linköping: Linköping University Electronic Press, 2020. 12-22.
- Frick, E., Helmer, H. & Wallner, F. (2023). *ZuRecht: Neue Recherchemöglichkeiten in Korpora gesprochener Sprache für Gesprächsanalyse und Deutsch als Fremd- und Zweitsprache*. Fandrych et al. (2023a), 44-71.
- Frick, E. & Schmidt, T. (2020). *Using full text indices for querying spoken language data*. Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora, Marseille, France, May. European Language Resources Association, 40-46.
- Frick, E. & Schmidt, T. (2025). *Querying spoken language data*. Bański, P., Heid, U. & Herzberg, L. (eds.): Standards for language data and infrastructures. Series 'Digital Linguistics'. De Gruyter.
- Frick, E., Schmidt, T. & Helmer, H. (2022). *Querying Interaction Structure: Approaches to Overlap in Spoken Language Corpora*. Proceedings of LREC 2022. Marseille, France. European Language Resources Association, 715-722.
- Frick, E., Lösel, A., Messner, A., Schmidt, T., Schwendemann, M. & Wallner, F. (2025). *ExpoKo – eine Korpusressource für das wissenschaftliche Schreiben*. Große Lernerkorpora - Möglichkeiten und Grenzen, Leipzig. Zenodo. <https://doi.org/10.5281/zenodo.17207106>
- Grän, J., Schaber, J., McDonald, D., Mustač, I., Rajović, N., Schneider, G., Vuković, T., Zehr, J. & Bubenhofer, N. (2024). *The LiRI Corpus Platform (LCP): A software system and infrastructure for querying a vast array of corpora of different kinds*. CLARIN Annual Conference 2023. <https://doi.org/10.3384/ecp210010>
- Hedeland, H. & Ferger, A. (2020). *Towards Continuous Quality Control for Spoken Language Corpora*. International Journal of Digital Curation 15. <https://doi.org/10.2218/IJDC.V15I1.601>
- Hedeland, H. & Schmidt, T. (2022). *The TEI-based ISO Standard 'Transcription of spoken language' as an Exchange Format within CLARIN and beyond*. Monachini, M. & Eskevich, M. (eds.): Selected Papers from the CLARIN Annual Conference 2021. Virtual Event, 2021, 27-29 September. (= Linköping Electronic Conference Proceedings 189). Linköping: Linköping University Electronic Press, 2022, 34-45.
- Kisler, T., Reichel, U. & Schiel, F. (2017). *Multilingual processing of speech via web services*. Computer Speech & Language, Volume 45, September 2017, pages 326-347.
- Körner, E., Eckart, T., Helfer, F. & Kretschmer, U. (2024). *An Enhanced Federated Content Search Infrastructure for the Humanities and Social Sciences*. CLARIN Annual Conference, 49-59. <https://doi.org/10.3384/ecp216.05>
- Krause, T. & Zeldes, A. (2016): ANNIS3: A new architecture for generic corpus query and

- visualization. In: *Digital Scholarship in the Humanities*, 31(1). 118-139.
<https://doi.org/10.1093/lle/fqu057>
- Kren, M. (2020): *Czech National Corpus in 2020: Recent Developments and Future Outlook*. In Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora, Marseille, France. European Language Resources Association, 52–57.
- Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (2017): *The Spoken BNC2014: designing and building a spoken corpus of everyday conversations*. In: *International Journal of Corpus Linguistics*, 22:3. 319-344. <https://doi.org/10.1075/ijcl.22.3.02lov>
- Monty Python (1970): *Royal Society For Putting Things On Top of Other Things*.
<https://www.youtube.com/watch?v=LFrDqQZ8FFc>
- Pęzik, P. (2015): *Spokes – a Search and Exploration Service for Conversational Corpus Data*. In Selected Papers from the CLARIN 2014 Conference. Linköping Electronic Conference Proceedings. Linköping University Electronic, 99–109.
- Pluschkovits, M., Wittibschlager, A., Schopper, D., Bal, J., Kukelka, K., Reichl, O., Rohler, M. & Lenz, A. (2026): *LAPIS: ein Portal zu Sprachressourcen in Österreich*. Contribution to the methods fair of the annual conference of the Institute for the German Language. https://www.ids-mannheim.de/fileadmin/aktuell/Jahrestagungen/2026/Methodenmesse/02_LAPIS.pdf
- Reichel, U.D., Kisler, T. (2014): *Language-independent grapheme-phoneme conversion and word stress assignment as a web service*. In: Hoffmann, R. (Ed.): *Elektronische Sprachverarbeitung. Studentexte zur Sprachkommunikation* 71, pp 42-49, TUDpress, Dresden.
- Sauer, S. & Lüdeling, A. (2016). *Flexible Multi-Layer Spoken Dialogue Corpora*. *International Journal of Corpus Linguistics* 21(3), Special Issue on Spoken Corpora, 419-438.
- 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., Paggio, P. & Heylen, D. (eds.): *Multimodal corpora: from models of natural interaction to systems and applications*. Berlin/Heidelberg: Springer, 2009, 207-221.
- Schmidt, T., Elenius, K. & Trilsbeek, P. (2010). *Multimedia Corpora (Media encoding and annotation)*. Zenodo. <https://doi.org/10.5281/zenodo.18269160>.
- Schmidt, T. (2011). *A TEI-based approach to standardising spoken language transcription*. *Journal of the Text Encoding Initiative* 1. 2011. <https://doi.org/10.4000/jtei.142>
- Schmidt, Thomas (2014a). *The Database for Spoken German – DGD2*. Proceedings of LREC 2014, Reykjavik, Iceland. European Language Resources Association (ELRA). 1451-1457.
- Schmidt, Thomas (2014b). *(More) Common Ground for Processing Spoken Language Corpora?* In: Ruhi, Ş., Haugh, M., Schmidt, T. & Wörner, K. (eds.): *Best Practices for Spoken Corpora in Linguistic Research*. Newcastle: Cambridge Scholars Publishing, 2014. S. 249-265.
- Schmidt, T. & Wörner, K. (2014). *EXMARaLDA*. Durand, J., Gut, U. & Kristoffersen, G. (eds.): *The Oxford Handbook of Corpus Phonology*, Oxford: OUP, 402-419.
- Schmidt, T. (2016). *Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project*. Kupietz, M. & Geyken, A. (eds.): *Corpus Linguistic Software Tools*. *Journal for language technology and computational linguistics (JLCL)* 31 (1). Berlin: GSCL, 2016. 127-154. <https://doi.org/10.21248/jlcl.31.2016.205>
- Schmidt, T., Hedeland, H. & Jettka, D. (2017). *Conversion and annotation web services for spoken language data in CLARIN*. In: Borin, L. (ed.): *Linköping Electronic Conference Proceedings; Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*. Linköping: Linköping University Electronic Press, 2017. S. 113-130.
- Schmidt, T., Schwendemann, M. & Wallner, F. (2023). *ZuViel: Transkriptvisualisierung und Arbeiten mit Transkripten*. Fandrych et al. (2023a), 72-91.
- Schmidt, T. & Frick, E. (2025). *Reading, listening to and watching concordances of audiovisual interaction corpora*. Poster presented at “Symposium: Reading Concordances in the 21st Century”, University of Erlangen, 19-21 March 2025. <https://doi.org/10.5281/zenodo.15094134>
- Schmidt, T. (2025). *Représenter et accéder à la parole dans les corpus oraux : diversification et adaptation des méthodes et technologies*. Kanaan-Caillol, L., Dugua, C. & Gerstenberg, A. (eds.), *Représenter la parole*. De Gruyter, 185-208.
- Schmidt, T. (2026, accepted). *Chapter 3: Typology of Non-Textual Language Data*. Witt, A., Lenz, A., Kamocki, P. (eds.): *Routledge Handbook of Digital Linguistics*. London: Routledge.

- Schmidt, T., Blevins, M., Boas, H. & Gilbert, G. (2026, accepted). *The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact*. Proceedings of LREC Workshop “Dialects in NLP: A Resource Perspective”.
- Sidnell, J. (2012). *Basic Conversation Analytic Methods*. In: Sidnell, S. & Stivers, T. (eds.): *The Handbook of Conversation Analysis*. Oxford: Blackwell, 77-99.
- Wamprechtshammer, A., Arestau, E., Aznar, J., Hedeland, H., Isard, A., Khait, I., Lange, H., Majka, N. & Rau, F. (2022). *QUEST. Guidelines and Specifications for the Assessment of Audiovisual, Annotatated Language Data*. Working Papers in Corpus Linguistics and Digital Technologies: Analyses and Methodology. Vol. 8 (2022). <https://ojs.bibl.u-szeged.hu/index.php/wpcl/issue/view/2474/382>.
- Westpfahl, S. & Schmidt, T. (2016). *FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German*. Proceedings of LREC 2016, Portoroz, Croatia. European Language Resources Association (ELRA), 2016. 1493-1499.
- Wörner, K. (2012). *Finding the balance between strict defaults and total openness - Collecting and managing metadata for spoken language corpora with the EXMARaLDA Corpus Manager*. In: Schmidt, T. & Wörner K. (eds.): *Multilingual Corpora and Multilingual Corpus Analysis*: 383–400, 18 S., Vol. 14, Amsterdam; Philadelphia: John Benjamins Pub. Co (Hamburg Studies on Multilingualism).