

The TEI-based ISO Standard ‘Transcription of spoken language’ as an Exchange Format within CLARIN and beyond

Hanna Hedeland

Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
hedeland@ids-mannheim.de

Thomas Schmidt

Research and Infrastructure Support
Universität Basel, Switzerland
th.schmidt@unibas.ch

Abstract

This paper describes the TEI-based ISO standard 24624:2016 ‘Transcription of spoken language’ and other formats used within CLARIN for spoken language resources. It assesses the current state of support for the standard and the interoperability between these formats and with relevant tools and services. The main idea behind the paper is that a digital infrastructure providing language resources and services to researchers should also allow the combined use of resources and/or services from different contexts. This requires syntactic and semantic interoperability. We propose a solution based on the ISO/TEI format and describe the necessary steps for this format to work as an exchange format with basic semantic interoperability for spoken language resources across the CLARIN infrastructure and beyond.

1 Introduction

Today, the CLARIN infrastructure is well established across Europe, comprising a network of centres providing a vast number of digital resources and services. Since an increasing number of funders require researchers in the humanities and social sciences to deposit their data for reuse, the collections of digital resources hosted within CLARIN are growing steadily. Following the digital turn, the use of CLARIN’s tools and services for manual and automatic analysis has also become a relevant option for research projects from various disciplines. An ideal scenario would allow researchers to use and freely combine data and tools or services from different CLARIN centres and contexts across the infrastructure. This, however, is still possible only for smaller sets of resources – large scale interoperability remains a desideratum. Unlike early digital corpora created by pioneering corpus linguists, digital language resources today seldom fit into the traditional view of language data as ‘natural running text’ or ‘a single stream of tokens’. This is particularly true for spoken or multi-modal resources, which are at the same time no longer a rare exception in the resource landscape.

In this paper, a TEI-based ISO standard for the representation of spoken language transcription will be introduced and its current and future relevance for CLARIN and related contexts will be discussed. After this introduction we will provide an overview of tools and services which are currently available to work with that standard in creating, enriching and publishing spoken language data.

2 A Standard for Spoken Language Transcription?

2.1 Interoperability of Existing De-Facto Standards and Tool Formats

One reason for the heterogeneity of spoken language corpora is the existence of several widely used tool formats. ELAN (Sloetjes, 2014), Praat (Boersma, 2001), CLAN (MacWhinney, 2000), Transcriber (Barras et al., 2001), FOLKER (Schmidt, 2016) and EXMARaLDA (Schmidt and Wörner, 2014) all come with their individual formats, which are, apart from Praat’s TextGrid format, XML-based. These formats are mainly based on similar tier-/time-based data models, i.e. they model transcription as a set of

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

tiers with different characteristics containing different information, and are already to a sufficient extent interoperable – from the syntactic perspective (Schmidt et al., 2009). A file in one format can usually be converted into a file with a representation of the data using another format. There are undoubtedly some limitations regarding conversion scenarios, depending on the varying complexity of data models, where e.g. certain tier hierarchies or associations between annotation elements in ELAN’s EAF format cannot be modelled by the more restrictive data model for Basic Transcriptions (EXB) in the EXMARaLDA system. In these rather rare cases, customised workarounds are still possible.

From a semantic perspective however, interoperability is not that straightforward, since both the set of tiers used and their content vary to a great extent. One solution to this dilemma would be to standardise tiers and tier content. As an example, the CHAT format of the CLAN software, depicted in Figure 1, exactly defines the set of transcription and annotation conventions to be used for common spoken language phenomena, which makes the data easy to process and understand. But researchers are at the same time required to subscribe to theoretical concepts implemented by these conventions, and this is not a good basis for a standard to be used across discipline boundaries.

```

@Begin
@Languages:      eng
@Participants:  CHI Ross Child, FAT Brian Father
@ID:            eng|macwhinney|CHI|2;10.10||||Target_Child|||
@ID:            eng|macwhinney|FAT|35;2.||||Target_Child|||
*ROS:          why isn't Mommy coming?
%com:          Mother usually picks Ross up around 4 PM.
*FAT:          don't worry.
*FAT:          she'll be here soon.
*CHI:          good.
@End

```

Figure 1: The CHAT transcription system defines the units of the transcription, the annotation tiers and the transcript layout.

On the other side of the spectrum, the EAF format of the ELAN software hardly imposes any restrictions on the individual researcher, who is free to define the structure and content of the data format according to her needs. While this promises a perfect fit for the individual research context, data modelling is not trivial and not all variation is semantically relevant. This means that transcripts containing e.g. a basic orthographic transcription, interlinear glosses and a translation into English can be modelled in various ways using different tier types and names, making automatic processing of similar resources difficult since the semantics of the tiers are only documented for humans. It should be noted that ELAN has been providing means to define the semantics of tiers and annotations using external controlled vocabularies or references to ISOcat for many years. The comprehensive evaluation of annotation practices in language documentation corpora presented by von Prince and Nordhoff (2020) shows that this has however hardly been adopted by researchers using the software. This might be related to the proliferation of data categories in ISOcat or simply a matter of lacking awareness of the problem.

2.2 The ISO/TEI Approach to Standardisation and Interoperability

The ISO standard for Transcription of spoken language (ISO/TC 37/SC 4, 2016; Schmidt, 2011) is based on the TEI Guidelines (TEI Consortium, 2021), mainly on the chapter ‘8 Transcriptions of Speech’¹. The idea behind the standard is to find a solution that differentiates between general information that is shared across different research methods and disciplines on the one hand, and information that is theory-dependent (cf. (Ochs, 1979)) and therefore cannot be standardised, on the other. Standardisation can be applied to aspects of the shared reality of spoken conversation, which includes e.g. the modelling of participants and the temporal alignment of their contributions. These aspects, referred to here as macro-structure, are not defined by transcription conventions or other theoretical constructs.

¹<https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html>

Unlike many of the widely used transcription tool formats, the ISO/TEI format depicted in Figure 2 is not a pure tier-/time-based format. Instead, it models speaker contributions as a common list of <u> elements. Its structure is thus more similar to written documents. Speaker contributions are often considered to comprise several linguistic units, accordingly <u> elements may contain one or more <seg> elements corresponding to the linguistic units defined by the relevant transcriptions system via @type and @subtype attributes. References to defined speakers and time points are modelled by the attributes @who, @start and @end, with the option to use <anchor> elements for additional alignment in any position². Annotations are by default modelled in a standoff manner by elements in <spanGrp> elements, the annotation level defined by a @type attribute. Annotations can be used with <anchor> elements as in Figure 2 or refer to words, <w> elements, if the text has been tokenised and marked-up accordingly. An additional element <annotationBlock> is used to group the speaker contribution <u> with all annotations referring to it.

```

<text xml:lang="en">
  <timeline unit="s">
    <when xml:id="T0" />
    <when xml:id="T3" />
    <when xml:id="T4" />
    <when xml:id="T1" />
    <when xml:id="T2" />
  </timeline>
  <body>
    <annotationBlock xml:id="AB1" who="#TIM" start="#T0" end="#T1">
      <u xml:id="U1">Excuse me, is <anchor synch="#T3" />this <anchor synch="#T4" />the way out? </u>
      <spanGrp type="pho">
        <span from="#T3" to="#T4">[zis]</span>
      </spanGrp>
    </annotationBlock>
    <annotationBlock xml:id="AB2" who="#TOM" start="#T1" end="#T2">
      <u xml:id="U2">Yes, straight ahead. </u>
    </annotationBlock>
  </body>
</text>

```

Figure 2: A simple example of the transcription macro-structure of the ISO/TEI format.

Below the macro-structure, within the speaker contribution, there are many differences in the precise form of representation for verbal and accompanying non-verbal elements and features across transcription systems. We will refer to this level, which may also contain widely recognised linguistic units such as words, as the micro-structure. The differences between the representations used in various transcription systems are partly due to important reflections of theoretical differences, but in other cases the syntactic differences resulting from the choices of transcription symbols do not reflect any semantic differences, and in some cases syntactic or symbolic identity obscures semantic differences. Figure 3 shows the traditional printed representation of the same speaker contribution using two different transcription systems.

MJ: I ((cough)) see a door. I (0.3) want to paint it (black/blue).

MJ[v] I ((cough)) see a door. I ((0,3s)) want to paint it (black).
MJ[k] (blue)

Figure 3: The same speaker contribution transcribed according to two different transcription systems; GAT (Selting et al., 1998) (above) and HIAT (Rehbein et al., 2004) (below)

To the human reader, the similarities are striking and the slight differences in the representation of identical phenomena are easily deciphered. Both transcription systems use double parentheses to represent non-verbal and non-phonological elements, the green highlighting of the ‘((cough))’ was therefore added to this example to indicate syntactic and semantic identity. The short (0.3 seconds) pause and the uncertainty regarding which colour (black or blue) will be used to paint the door share the same semantics

²Owing to performance reasons and ease of processing, the ZuMult project (cf. Section 4.3) uses ID/IDREFs instead of XPointers for pointing between elements.

but are syntactically different, the added highlighting is yellow. The uncertain part is even structurally different, since the HIAT system (below in Figure 3) requires the alternative interpretation to be transcribed in an additional tier for comments ('k') below the main transcription tier. The two full stops highlighted in red are on the contrary syntactically identical, but their semantics differ, since the two transcription systems use this symbol to denote different types of units within the speaker contribution.

It is possible to represent this example in the ISO/TEI format without taking the transcription conventions into account. In Figure 4 this has been done for the same example with the GAT version above and the HIAT version below. The same similarities and differences still apply and the structural difference in the representation of uncertainty is encoded once through symbols in the text of the speaker contribution for GAT (above in Figure 4) and once as an annotation of the uncertain part for HIAT (below in Figure 4). With this representation of the data in the same format, syntactic interoperability has been achieved. Reliable automatic processing or querying of the content of this type of data across collections using different transcription systems still remains difficult, since there is no semantic interoperability on this level.

```

<u who="#MJ" start="#T0" end="#T2">
  <seg type="contribution">
    I ((cough)) see a door.
    <anchor synch="#T1"/>
    I (0.3) want to paint it (black/blue).
  </seg>
</u>

<u who="#MJ" start="#T0" end="#T4">
  <seg type="contribution">
    I ((cough)) see a door.
    <anchor synch="#T1"/>
    I ((0,3s)) want to paint it <anchor synch="#T2"/>(black)<anchor synch="#T3"/>.
  </seg>
</u>
<spanGrp type="k" subtype="time-based">
  <span from="#T2" to="#T3">(blue)</span>
</spanGrp>

```

Figure 4: The examples can be represented in the ISO/TEI format without using the implicit information of the transcription conventions.

```

<u who="#MJ" start="#T0" end="#T2">
  <seg type="intonation-phrase" subtype="falling">
    <w>I</w><vocal><desc>cough</desc></vocal><w>see</w><w>a</w><w>door</w>
  </seg>
  <anchor synch="#T1"/>
  <seg type="intonation-phrase" subtype="falling">
    <w>I</w><pause dur="PT0.3S"/><w>want</w><w>to</w><w>paint</w><w>it</w>
    <unclear><choice><seg><w>black</w></seg><seg><w>blue</w></seg></choice></unclear>
  </seg>
</u>

<u who="#MJ" start="#T0" end="#T2">
  <seg type="utterance" subtype="declarative">
    <w>I</w><vocal><desc>cough</desc></vocal><w>see</w><w>a</w><w>door</w>
  </seg>
  <anchor synch="#T1"/>
  <seg type="utterance" subtype="declarative">
    <w>I</w><pause dur="PT0.3S"/><w>want</w><w>to</w><w>paint</w><w>it</w>
    <unclear><choice><seg><w>black</w></seg><seg><w>blue</w></seg></choice></unclear>
  </seg>
</u>

```

Figure 5: When encoded using the ISO/TEI format, the partly identical meaning of the different transcription symbols becomes explicit and only the theory-dependent differences of Figure 4 remain.

Semantic interoperability can be achieved through standardisation, though while some aspects of the micro-structure can be standardised, such as the existence of pauses and (possibly) non-verbal behaviour, the detailed choices regarding e.g. a set of relevant different pause durations or the descriptions of non-verbal behaviour have to correspond to the theory-dependent transcription system currently in use. The same is true for the details of the segmentation into linguistic units in <seg>s, which usually differs according to the linguistic level used as the basis. Allowing for controlled variation within this area makes it possible to represent data created with different transcription systems using the same standard format. In Figure 5 the micro-structure has been parsed according to the different transcription systems during the conversion process and a common representation of shared phenomena – the word and non-word tokens, the pause, and the uncertainty with the alternative interpretation – has been achieved. It has also become possible to explicitly express the different semantics of the units below the speaker contribution, i.e. the different meaning of the full stop in the two transcription systems, through the use of <seg> elements with @type and @subtype attributes, in this case intonation phrases based on interactional prosody for the GAT system and utterances based on the pragmatics level for the HIAT system. This type of conversion results in transcription data that is semantically interoperable where this is possible and for which semantic and theory dependent differences become explicit and machine-readable.

3 Acceptance of ISO/TEI and Related Formats in CLARIN

Within CLARIN, centres are not bound to accept or support particular formats. In accordance with the requirements of the CoreTrustSeal (CoreTrustSeal Standards and Certification Board, 2019), which is a prerequisite for the certification of CLARIN B centres (cf. (Wittenburg et al., 2019)), all centres do however provide information about accepted file formats for resource deposits. Some centres have compiled individual lists for this purpose and others still refer to one of several older general lists and overviews of standards and recommendations for CLARIN³. While these lists pre-date the ISO/TEI format, they all include TEI as a general recommendation. At the time of writing, seven B centres point to such external information⁴.

The CLARIN Standards Committee has been gathering information on the recommendations on standards and formats actively issued by individual (mainly B) centres and made this information available on their web page⁵ and as the basis for the relaunch of the CLARIN Standards Information System (SIS)⁶. A brief assessment of this information can provide insights into the current and potential support for the ISO/TEI standard within CLARIN. For this paper, the Standards Information System and the original recommendations given by individual centres were surveyed. Since the transformation from the various centres' individual recommendations into the SIS might be a source of inaccuracy, the original documents and websites were revisited for centres that have not validated and confirmed their SIS information. As not all centres accepting data deposits provide detailed individual recommendations yet, the picture is however still not complete. Since there is also no consistent and reliable information on the general types of resources a centre accepts nor on specific restrictions e.g. regarding languages or time periods, negative results cannot really be interpreted in the sense of lacking acceptance for ISO/TEI or related formats, since the centre might not accept resource types for which ISO/TEI is a relevant format.

Nevertheless, of the centres that provide their own preferences and recommendations, three groups with respect to ISO/TEI support can be distinguished. According to validated information of the SIS and the centres' original recommendations at the time of writing, four B centres already recommend ISO/TEI explicitly⁷. These are the CLARIN.SI Language Technology Centre, The Language Bank of Finland (FIN-CLARIN), the Hamburg Centre for Language Corpora (HZSK) and the Leibniz-Institut für Deutsche Sprache (IDS). In addition to the information from certified B centres, the centres TOols for LANGuage (ORTOLANG) and Language Archive Cologne (LAC), which are both participating in

³Such resources are e.g. <https://www.clarin.eu/faq/what-standards-are-recommended-clarin> or <https://www.clarin.eu/sites/default/files/Standards\%20for\%20LRT-v6.pdf>

⁴cf. <https://github.com/clarin-eric/standards/issues/14>

⁵<https://www.clarin.eu/content/standards>


⁶<https://standards.clarin.eu/sis/>

⁷cf. <https://standards.clarin.eu/sis/views/view-format.xq?id=fTEISpoken>

ISO/TEI Transcriptions of Spoken Language

Abbreviation: TEISpoken

Identifiers:

Type	Id
SIS ID	fTEISpoken 

Media type(s):

- application/tei+xml;format-variant=tei-iso-spoken
- application/tei+xml;format-variant=tei-iso-spoken;tokenized=[0,1]

File extension(s): .tei

Format family: TEI

Recommendation:

Clarín Centre	Domain	Level	Comments
HZSK	Audiovisual Annotation	recommended	
ZIM	Audiovisual Annotation	recommended	
IDS	Audiovisual Annotation	recommended	
FIN-CLARIN	Audiovisual Annotation	recommended	
CLARIN.SI	Audiovisual Annotation	recommended	

Description:

This format is a serialization of the [ISO/TEI Transcriptions of Spoken Language](#).

ISO/TEI transcriptions of spoken language will be identified by the MIME type `application/tei+xml;format-variant=tei-iso-spoken`. A parameter `tokenized=0/1` can be added to indicate whether (=1) or not (=0) the respective TEI file is tokenized (i.e. has `<w>` markup).

For more information, see [Thomas Schmidt, "A TEI-based Approach to Standardising Spoken Language Transcription", Journal of the Text Encoding Initiative \[Online\], Issue 1 | June 2011, Online since 08 June 2011, connection on 21 September 2021. URL: <http://journals.openedition.org/jtei/142>; DOI: <https://doi.org/10.4000/jtei.142>](#)

Please feel welcome to supply the description of this format file via GitHub: either as an [issue report](#), or as a [pull request](#) after forking or browsing the [code](#) under the 'formats' branch.

[\[suggest a fix or extension\]](#)

Keywords: annotation format, corpus encoding

Figure 6: Information on the ISO/TEI format in the CLARIN Standards Information System (SIS).

CLARIN knowledge centres and aiming for B Centre status, also explicitly recommend the ISO/TEI format for incoming deposits.

Forming a second group, further centres recommend TEI, and thus implicitly ISO/TEI, though this variant is not explicitly mentioned⁸. Among these are the Austrian Centre for Digital Humanities and Cultural Heritage - A Resource Centre for the HumanitiEs (ACDH-ARCHE), Eberhard Karls Universität Tübingen (EKUT), the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), The CLARIN Centre at the University of Copenhagen (CLARIN-DK-UCPH), the ZIM Centre for Information Modelling (ZIM) and the Meertens Instituut/HuC (MI) (which only includes XML in the list, but refers to TEI as an example). The centre Collections de corpus oraux numériques (COCOON) also recommends TEI, is however not a certified B centre. As noted above, all centres referring to existing CLARIN documents also in effect recommend TEI without further restrictions.

The third group is the most interesting, since these centres explicitly recommend other widely used formats and not ISO/TEI. The CMU-TalkBank (CMU) recommends CHAT (only), MPI for Psycholinguistics (MPI-PL) recommends CHAT too, though in addition to EAF and Praat, which are in turn also

⁸cf. <https://standards.clarin.eu/sis/views/view-format.xq?id=fTEI>

recommended by the Bayerisches Archiv für Sprachsignale (BAS). Both Praat and EAF can be converted into the ISO/TEI format with dedicated software as described in (Schmidt et al., 2017), and this also applies to CHAT data that passes the data quality and consistency tests in CLAN. Still, the ISO/TEI format seems to be of little relevance to these four centres, presumably because of strong traditions and eco-systems around specific formats for specific types of resources and research areas. Furthermore, the LINDAT/CLARIAH-CZ centre, which does not give explicit recommendations on formats to depositors, now hosts the TEI-based TEITOK system (Janssen, 2021), which includes both a search engine, visualisation and editing functionality and has many features for spoken language. Since this TEI variant is interoperable with e.g. EXMARaLDA and EAF through a set of scripts, interoperability between the TEITOK and ISO/TEI formats is also feasible.

As expected, TEI, the ISO/TEI format, and formats that can be converted into the ISO/TEI formats are often recommended for resource deposition across the infrastructure. A more systematic approach towards the description and dissemination of format recommendations would facilitate further steps towards enhanced interoperability for transcription data in CLARIN. The Standards Information System can now be used to manage and analyse the relevant information as provided by the centres.

4 Tools and Services for ISO/TEI within and beyond CLARIN

Whether or not a new standard is widely adopted crucially depends on how well it interoperates with existing tools and methods. Ideally, researchers can continue working with established workflows and will profit from additional benefits because these workflows are becoming standard-compliant. The ISO/TEI standard was defined with this practical goal in mind. In what follows we will look at different stages of the research data lifecycle for spoken language corpora, explaining and illustrating how existing tools and methods interoperate with the standard.

4.1 Data Creation (Transcription)

Among the existing, widely used tools for transcription (see above), the EXMARaLDA Partitur-Editor and FOLKER/OrthoNormal provide the most direct interoperability with ISO/TEI. The tools continue to write their tool specific format, but now have an additional option for exporting ISO/TEI. In the case of the Partitur-Editor, the export can be configured to use different algorithms for segmenting transcribed text into word and non-word tokens (such as pauses or descriptions of non-verbal behaviour) according to different transcription systems (see Figure 7). The Partitur-Editor can also import files in the ISO/TEI format. Since the internal tool format does not represent tokens and other parts of the micro-structure, this is strictly speaking a lossy transformation. The information, however, can be automatically reconstructed from implicit information during the corresponding export process.

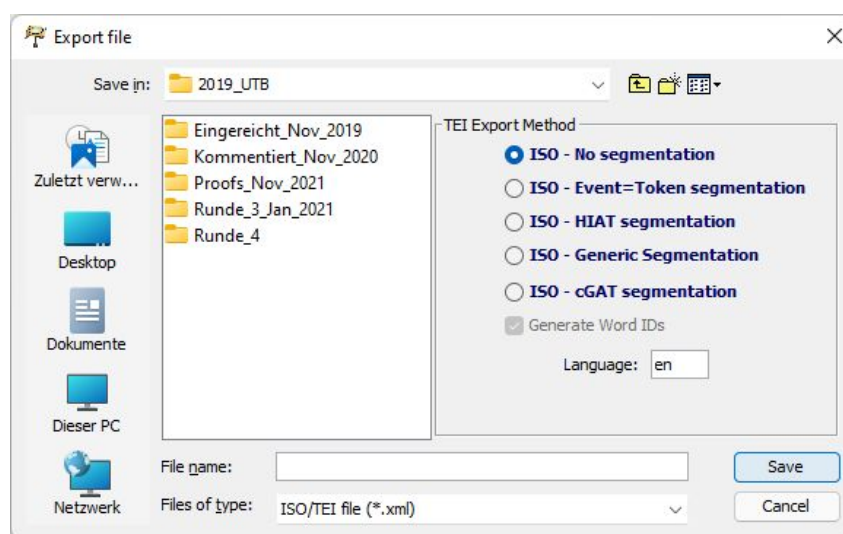


Figure 7: ISO/TEI export dialog of the EXMARaLDA Partitur-Editor.

The other transcription tools mentioned above (i.e. ELAN, Transcriber, CLAN and Praat) do not (as yet) provide direct means of importing or exporting ISO/TEI. The conversion can, however, be achieved via the EXMARaLDA Partitur-Editor (which has import filters for all of the formats), via TEI-Drop, a dedicated tool for that purpose, or via web-services (Schmidt et al., 2017).

4.2 Data Enrichment (Annotation)

Since the creation of the ISO/TEI standard, the format has been used as the basis for enhanced interoperability with existing annotation tools and services. In many cases, this was software created on the basis of data models or notions of written language. Since the ISO/TEI standard is a TEI-based format, it shares a common core with TEI variants used for written language data and thus facilitates interoperability across the spoken and written modality. For instance, the development of WebAnno-MM (Remus et al., 2019) as an extension for audiovisual and transcription data in the ISO/TEI format allows manual annotation with a wider textual focus than transcription tools offer, and also more complex types of annotations such as tree or chain annotations. The original user interface for annotation tasks and the score visualisation for transcription data are shown in Figure 8.

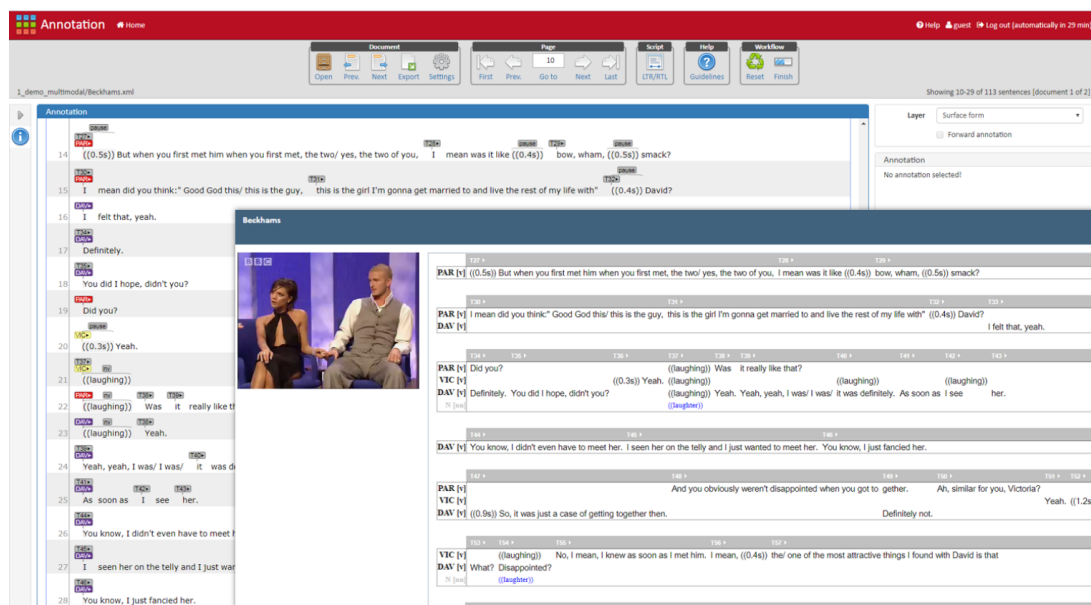


Figure 8: Annotation and multimedia transcript score view in WebAnno-MM.

For automatic annotation, the converters described above were integrated into the WebLicht SOA (Hinrichs et al., 2010) of CLARIN-D, thus enabling the use of various services from all German centres. Initially, this meant another mapping to formats and services for written data (internally, TCF, see

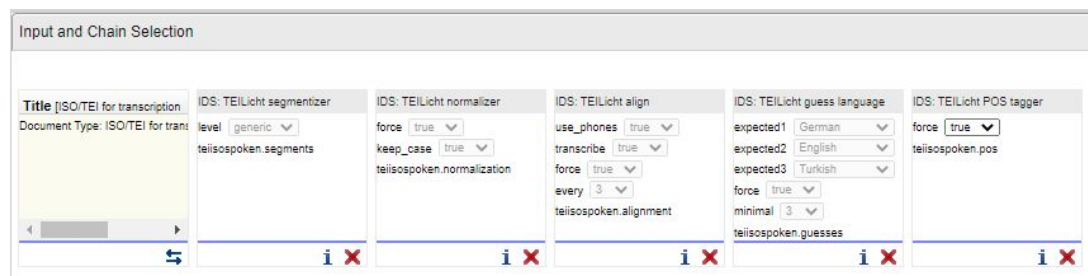


Figure 9: An ISO/TEI annotation chain defined in WebLicht.

(Schmidt et al., 2017)), but services adapted to spoken language data based directly on the ISO/TEI format have now also been developed (Fisseni and Schmidt, 2020) and can improve results where the

linguistic characteristics of spoken and written language differ to a great extent. A sample processing chain is shown in Figure 9. The speech data web services provided by the BAS (Kisler et al., 2017) have been able to import and export ISO/TEI data since version 2.36 of January 2020.

4.3 Data Publication and Analysis (Dissemination)

Based on the ISO/TEI format, the project ZuMult has developed new web-based functionality for both visualisation and browsing of spoken language corpora within qualitative approaches and for complex querying and analysis⁹. Query is based on an extension of the MTAS system (Brouwer et al., 2017) which can generate Lucene indices directly from the ISO/TEI XML files. Users can thus be provided with very powerful and efficient querying possibilities in CQP (Frick and Schmidt, 2020). Visualisation uses various XSL transformations to generate, directly from the ISO/TEI XML file, configurable displays of the transcript (in HTML), a density viewer (in SVG) and configurable video subtitles (in VTT) all of which are synchronised with each other and with the underlying audio or video (see Figure 10).

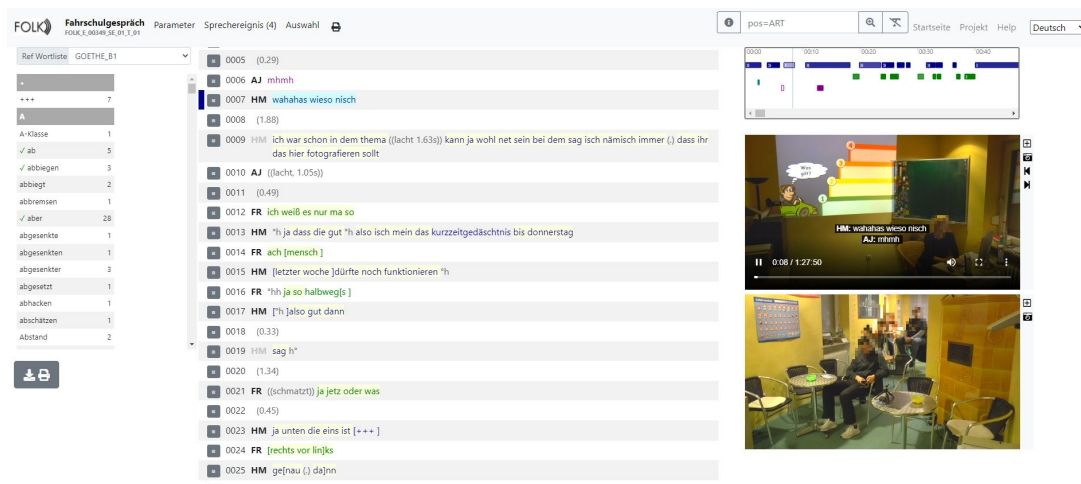


Figure 10: Different visualisations of an ISO/TEI transcript, integrated and synchronised in the ZuViel tool of the ZuMult project.

Another corpus analysis platform that now supports the ISO/TEI format is Tsakorpus (Arkhangelskiy et al., 2019), which is one use case for ISO/TEI within the long-term project INEL in Hamburg (Ferber and Jettka, 2020). A project in the related field of language documentation, the international (French/German) DoReCo project (Paschen et al., 2020), developed the Multitool¹⁰ that can generate ISO/TEI as a distribution format for resources in various languages and tool formats. The ISO/TEI standard is also used as a pivot format for different tool formats in the tool TEICORPO¹¹ developed at the CLARIN K centre CORLI to facilitate data sharing and long-term preservation (Parsse et al., 2020). Since the main aim is a direct lossless conversion from the ELAN, Praat, Transcriber and CHAT formats, the work is complementary to the existing solutions based on the EXMARALDA system. The conversion solutions developed at CORLI also focus on the macro-structure and TEI-conform means of representing arbitrary tier structures found in tool formats of varying complexity without attempts to map micro-structure information systematically.

5 Discussion

The development of interfaces between the ISO/TEI standard and various existing tools and services has shown that this is not only feasible, but also efficient using the ISO/TEI standard as a pivot format. This is important since software development and maintenance is usually the bottleneck in the development

⁹<http://zumult.ids-mannheim.de/ProtoZumult/index.jsp>

¹⁰<https://github.com/DoReCo/multitool>

¹¹<https://ct3.ortolang.fr/teicorpo/>

of the infrastructure. As (Parisse et al., 2020) point out, researchers also need to continue using tools they are familiar with. The ISO/TEI format could enhance interoperability for spoken language resources in CLARIN, especially since the already mentioned centres CORLI, LINDAT and IDS, and parts of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD), already actively use TEI for spoken data. Using a TEI variant to achieve interoperability has also proven successful in the case of parliament corpora (Erjavec et al., 2022). By using a TEI-based format for spoken data, apart from the proximity to more familiar written language data models on the textual level, interoperability on the metadata level could also be facilitated. With the TEI header, there is also a common structure for a core set of relevant contextual information on the setting and the participants, e.g. for analyses within virtual collections. Since TEI is used and extended in many contexts, there are also existing conventions for basic token-based linguistic annotation (Bański et al., 2018) and a common approach for the integration of the W3C standard RDFa is being developed (Chiarcos and Ionov, 2019) to tackle the issue of strict linked data requirements, which are also relevant for the interoperability aspects of the FAIR principles (Wilkinson et al., 2016).

Though conversion is already possible for widely used tool formats, as pointed out above, only features of the macro-structure are strictly defined by the ISO/TEI standard, and only syntactic interoperability is to some extent simple to achieve. For semantic interoperability, the tier structure, the annotation levels and schemas and the conventions for transcription – the micro-structure – also need to be made explicit and machine processable to allow for tokenisation and structural mark-up. This means that a conversion into the ISO/TEI format is not only a question of interoperability with a standard, but at the same time a process of FAIRification, of defining the semantic model of the data, making it more transparent and increasing the number and types of possible re-use scenarios. Creating digital language resources that are FAIR according to the well-known principles is a great, and often somewhat abstract, challenge for CLARIN and its users. We suggest that the adoption of the ISO/TEI standard with its basic semantics and the corresponding conversion scenarios as a way of assessing digital language resources could not only improve interoperability across resources, but also increase their general FAIRness. By using TEI as a common format and settling for answers to the question of machine-readable annotation documentation (Chiarcos et al., 2020) CLARIN could help foster a culture of data documentation required for interoperable and truly FAIR infrastructures for both humans and machines.

6 Conclusion

As this paper has tried to demonstrate, TEI-based standardisation for a sufficiently well-specified domain can make a contribution towards improved syntactic and semantic interoperability in a landscape where different tool-specific formats are already established. Although many issues still remain to be solved, we think that this approach is the most concrete and pragmatic that can be realised in a heterogeneous context such as CLARIN. The ISO/TEI standard, in this sense, is both a technical basis for data exchange in the ‘real world’ and a conceptual model for thinking about farther-reaching standardisation. Adopting such standard proposals as preferred formats of CLARIN centres can further help to consolidate such common ground.

References

- Timofey Arkhangelskiy, Anne Ferger, and Hanna Hedeland. 2019. Uralic multimedia corpora: ISO/TEI corpus data in the project INEL. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 115–124, Tartu, Estonia, January. Association for Computational Linguistics.
- Piotr Bański, Susanne Haaf, and Martin Mueller. 2018. Lightweight grammatical annotation in the TEI: New perspectives. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan*, pages 1795–1802, Paris, France. European language resources association (ELRA).
- Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22. Speech Annotation and Corpus Tools.

- Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glott International*, 5(9/10):341–345.
- Mathijs Brouwer, Hennie Brugman, and Marc Kemps-Snijders. 2017. MTAS: a solr/lucene based multi tier an-notation search solution. In *Selected papers from the CLARIN Annual Conference*, pages 19–37, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.
- Christian Chiarcos and Max Ionov. 2019. Linking the TEI: Approaches, Limitations, Use Cases. In *Digital Humanities Conference 2019 (DH2019)*, Utrecht University, July.
- Christian Chiarcos, Christian Fäth, and Frank Abromeit. 2020. Annotation Interoperability for the Post-ISOcat Era. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5668–5677, Marseille, France, May. European Language Resources Association.
- CoreTrustSeal Standards and Certification Board. 2019. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022, November.
- Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.
- Anne Ferger and Daniel Jettka. 2020. Use cases of the ISO standard for Transcription of spoken language in the project INEL. In *Proceedings of the CLARIN Annual Conference 2020*. CLARIN ERIC.
- Bernhard Fisseni and Thomas Schmidt. 2020. CLARIN web services for TEI-annotated transcripts of spoken language. Selected Papers from the CLARIN Annual Conference 2019. Leipzig, 30 September–2 October 2019, pages 12–22. Linköping University Electronic Press, Linköping.
- Elena Frick and Thomas Schmidt. 2020. Using full text indices for querying spoken language data. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 40–46, Marseille, France, May. European Language Resources Association.
- Erhard W. Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.
- ISO/TC 37/SC 4. 2016. Language resource management – Transcription of spoken language. Standard ISO 24624:2016, International Organization for Standardization, Geneva, Switzerland.
- Maarten Janssen. 2021. A corpus with wavesurfer and TEI: Speech and video in TEITOK. In Kamil Ekštejn, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue*, pages 261–268, Cham. Springer International Publishing.
- Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.
- Brian MacWhinney. 2000. *The CHILDES project: Tools for Analyzing Talk, Third edition. Volume I*. Lawrence Erlbaum, Mahwah, NJ u.a., 3rd edition.
- Elinor Ochs. 1979. Transcription as theory. In E. Ochs and B.B. Schieffelin, editors, *Developmental pragmatics*, pages 43–72. Academic Press, New York.
- Christophe Parrisé, Carole Etienne, and Loïc Liégeois. 2020. TEICORPO: a conversion tool for spoken language transcription with a pivot file in TEI. *Journal of the Text Encoding Initiative*, 13, May.
- Ludger Paschen, François Delafontaine, Christoph Draxler, Susanne Fuchs, Matthew Stave, and Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France, May. European Language Resources Association.
- Jochen Rehbein, Thomas Schmidt, Bernd Meyer, Franziska Watzke, and Annette Herkenrath. 2004. Handbuch für das computergestützte Transkribieren nach HIAT. *Arbeiten zur Mehrsprachigkeit, Folge B*, 56:1 ff.
- Steffen Remus, Hanna Hedeland, Anne Ferger, Kristin Bührig, and Chris Biemann. 2019. WebAnno-MM: EXMARALDA meets WebAnno. In *Selected papers from the CLARIN Annual Conference*, Pisa. Linköping University Electronic Press, Linköpings Universitet.

- Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.
- Thomas Schmidt, Susan Duncan, Oliver Ehmer, Jeffrey Hoyt, Michael Kipp, Dan Loehr, Magnus Magnusson, Travis Rose, and Han Sloetjes. 2009. An exchange format for multimodal annotations. In Michael Kipp, editor, *Multimodal corpora. From models of natural interaction to systems and applications*, Multimodal corpora. From models of natural interaction to systems and applications, pages 207 – 221. Springer, Berlin [u.a.].
- Thomas Schmidt, Hanna Hedeland, and Daniel Jettka. 2017. Conversion and annotation web services for spoken language data in clarin. In *Selected papers from the CLARIN Annual Conference*, pages 113–130, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.
- Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1, 06.
- Thomas Schmidt. 2016. Construction and dissemination of a corpus of spoken interaction - tools and workflows in the folk project. *Journal for language technology and computational linguistics (JLCL)*, 31(1):127 – 154.
- Margret Selting, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne Günthner, Christoph Meier, Uta Quasthoff, Peter Schlobinski, and Susanne Uhmann. 1998. Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, 173:91–122.
- Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.
- TEI Consortium. 2021. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Technical Report 4.3.0, TEI Consortium, August.
- Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2778–2787, Marseille, France, May. European Language Resources Association.
- Mark D. Wilkinson, 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, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.
- Peter Wittenburg, Dieter van Uytvanck, Thomas Zastrow, Pavel Straňák, Daan Broeder, Florian Schiel, Volker Boehlke, Uwe Reichel, and Lene Offersgaard. 2019. CLARIN B Centre Checklist (CE-2013-0095), Version 7.3.1, 2019-09-30. Technical report, CLARIN ERIC, September.