Originally published in: Monachini, Monica/Eskevich, Maria (Eds.): Proceedings of CLARIN Annual Conference 2021. 27 – 29 September 2021, Virtual Edition. – Utrecht: CLARIN, 2021. Pp. 100-104.

# 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 2462:2016 "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.

## 2 A standard for spoken language transcription?

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) 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 and to a sufficient extent interoperable – from the syntactic perspective. 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, but in these rather rare cases, customized workarounds are still possible.

From a semantic perspective however, interoperability is not that straightforward. As an example, the CHAT format of the CLAN software 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 disciplinary boundaries. 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. It should be noted that ELAN provides means for defining the semantics of tiers and annotations using references to ISOcat, but this has hardly been adopted as a practice by researchers (cf. von Prince and Nordhoff (2020)) and ISOcat had its own issues.

The idea behind the ISO standard for Transcription of spoken language (ISO/TC 37/SC 4, 2016; Schmidt, 2011) is a solution which 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 standardized, on the other. Standardization 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, referred to here as macro-structure. The ISO/TEI format is not a tier-/time-based format, but instead models speaker contributions as a common list of `<u>` elements, possibly containing one or more `<seg>` element for the linguistic units defined by the relevant transcriptions system via `@type` and `@subtype` attributes, e.g. `@type="intonation-phrase" @subtype="falling"`. Annotations are by default modelled by `<span>`s in `<spanGrp>`s with an additional element `<annotationBlock>` to group the speaker contribution `<u>` with all relevant annotations. References to defined speakers and time points are modelled by the attributes `@who`, `@start` and `@end`, with the option to use `<anchor>`s for additional alignment in any position.

While some aspects of speaker contributions can be standardized, 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 are not part of the standard but of the 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, e.g. intonation phrases for interactional prosody or utterances for pragmatics. Allowing for controlled variation within this area, the micro-structure, which defines the precise form of representation for spoken material, makes it possible to represent data created with different transcription systems using the same standard format.

## 3 Support for the ISO/TEI format in CLARIN

Within CLARIN, centres are not bound to accept or support particular formats, but several lists and overviews of standards and recommendations have been available over the years. Many centres refer to these resources[1] to define the formats they accept as deposits, e.g. for the Core Trust Seal, and thus include TEI as a general recommendation without further specifying any specific variants. 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 page[2]. 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 recommendations given by individual centres were revisited to allow for a more detailed picture. As not all (B) centres provide this information yet, the picture is however 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.

Nevertheless, of the centres that provide their own preferences and recommendations, three groups

[1] Such 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`

[2] `https://www.clarin.eu/content/standards`, this abstract is based on the Release 0.1 from January 2021 (the list at `https://www.clarin.eu/content/standards-and-formats` includes links to centres' original published documents in English)

with respect to ISO/TEI support can be distinguished. Three B centres already recommend ISO/TEI explicitly: the CLARIN.SI Language Technology Centre, the Hamburg Centre for Language Corpora (HZSK) and the Leibniz-Institut für Deutsche Sprache (IDS). The second group recommends TEI, but not explicitly ISO/TEI (or other variants). 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), Meertens Instituut/HuC (MI) (which only includes XML in the list, but refers to TEI as an example). And 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 recommended by The Language Bank of Finland (FIN-CLARIN) and 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. On the other hand, in addition to the information from certified B centres, there is information on accepted and recommended formats from the centres Open Resources and TOols for LANGuage (ORTOLANG) and Language Archive Cologne (LAC), which are both participating in knowledge centres and aiming for B Centre status: both recommend the ISO/TEI format for deposits. Furthermore, the LINDAT/CLARIAH-CZ centre, which does not give explicit recommendations on formats to depositors, now hosts the TEI-based TEITOK system (Janssen, 2016; Janssen, 2021), which includes both a search engine, visualization and editing functionality and has many features for spoken language. Since it is interoperable with e.g. EXMARaLDA and EAF through a set of scripts, interoperability between the TEITOK and ISO/TEI formats should not be difficult to establish.

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

For a standard to be useful to researchers and operators of research infrastructures, there need to be sufficient relevant use cases and software solutions that are compatible with existing tools and methods. For data creation, thanks to the existing conversion functionality described above (Schmidt et al., 2017), widely established tools can continue to be used. The EXMARaLDA transcription and annotation editor can not only export the ISO/TEI format, but also import these files e.g. after further enrichment outside of the EXMARaLDA environment.

Since the creation of the ISO/TEI standard, the format has been used as the basis for enhanced interoperability with several existing 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.

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 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. The phonetics 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.

Based on the ISO/TEI format, the project ZuMult has developed new web-based functionality for both visualization and browsing of spoken language corpora within qualitative approaches and for complex querying and analysis[3] based on an extension of the MTAS system (Brouwer et al., 2017) using CQP and a highly efficient query engine (Frick and Schmidt, 2020). 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 (Arkhipov and Däbritz, 2018; Ferger and Jettka, 2020). Another project in the field of language documentation, the DoReCo project (Paschen et al., 2020), developed the Multitool to generate ISO/TEI as a distribution format for resources in various tool formats. The use of the ISO/TEI standard as a pivot format for various language resources and different tool formats has also been implemented as a proof-of-concept workflow (Parisse et al., 2018).

## 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 of the infrastructure. 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.

Though conversion is already possible for widely used tool formats, as pointed out above, only features of the macro-structure are 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 tokenization 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 (Wilkinson and others, 2016) 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 and help foster a culture of data documentation required for truly FAIR infrastructures for both humans and machines.

## 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.

Alexander Arkhipov and Chris Lasse Däbritz. 2018. Hamburg corpora for indigenous Northern Eurasian languages. *Tomsk Journal of Linguistics and Anthropology*, (3):9–18.

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).

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Mathhijs 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.

[3] http://zumult.ids-mannheim.de/ProtoZumult/index.jsp

Christian Chiarcos and Max Ionov. 2019. Linking the TEI: Approaches, Limitations, Use Cases. In *Digital Humanities Conference 2019 (DH2019)*, Utrecht University, July.

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 Ressources 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 2462:2016, International Organization for Standardization, Geneva, Switzerland.

Maarten Janssen. 2016. TEITOK: text-faithful annotated corpora. In Nicoletta Calzolari et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23–28, 2016*. European Language Resources Association (ELRA).

Maarten Janssen. 2021. A corpus with wavesurfer and TEI: Speech and video in TEITOK. In Kamil Ekštein, 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 Parisse, Céline Poudat, Ciara R. Wigham, Michel Jacobson, and Loïc Liégeois. 2018. CORLI: A linguistic consortium for corpus, language, and interaction. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017*, pages 15–24, Budapest, Hungary. Linköping University Electronic Press, Linköpings Universitet.

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.

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, 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.

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.

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 et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.