Conversion and Annotation Web Services for Spoken Language Data in CLARIN

Thomas Schmidt
Institute for the German Language (IDS)
Mannheim
Germany
thomas.schmidt@ids-mannheim.de

Hanna Hedeland
Hamburg Centre for Language Corpora (HZSK)
University of Hamburg
Germany
hanna.hedeland@uni-hamburg.de

Daniel Jettka
Hamburg Centre for Language Corpora (HZSK)
University of Hamburg
Germany
daniel.jettka@uni-hamburg.de

Abstract

We present an approach to making existing CLARIN web services usable for spoken language transcriptions. Our approach is based on a new TEI-based ISO standard for such transcriptions. We show how existing tool formats can be transformed to this standard, how an encoder/decoder pair for the TCF format enables users to feed this type of data through a WebLicht tool chain, and why and how web services operating directly on the standard format would be useful.

1 Introduction

Web services operating on language resources are a central idea for CLARIN. This includes services for the annotation of text data, such as lemmatizers, POS-Taggers, Named Entity Recognizers, etc. The integration of services into the service-oriented architecture WebLicht (Hinrichs et al. 2010) enables users to apply customized annotation workflows to a set of data. So far, most such services were built with, and are meant to operate on, "canonical" written language data, typically edited texts from newspapers, books, etc., in the standard orthography of a major language. For "non-canonical" types of written data (such as computer-mediated communication (CMC) or historical language data) and for spoken language data, these services are often not directly usable because a) these data come in formats which are more complex than the simple "stream of tokens" (Menke et al. 2015) which many annotation services expect and/or b) the nature of the text data (e.g. non-standardized writing, lack of punctuation, unknown lexemes) requires additional processing steps or adaptations of annotation methods in order to yield useful results.

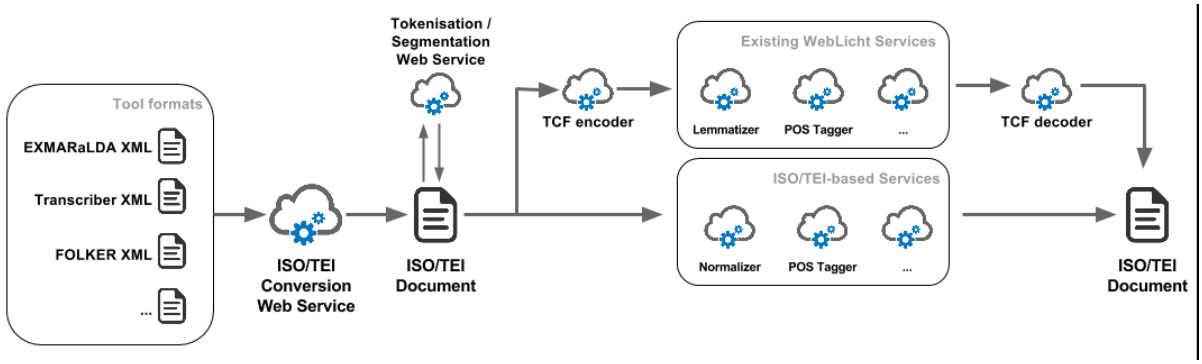

Figure 1. Architecture

We present an approach to making existing service environments and services in CLARIN usable for spoken language data. That this is possible and useful in principle has been demonstrated by proof-of-concept implementations in the tools EXMARaLDA (Schmidt/Wörner 2014) and ELAN (Kisler et al. 2012) which both provide interfaces to WebLicht. Our approach is more flexible because it is based on a TEI-based ISO standard for spoken language transcriptions and thus not directly tied to a specific tool or tool format. The general architecture we have started to implement is depicted in figure 1. The point of departure for most users will be one of a few established formats of tools for multimedia annotation. This needs to be converted to the ISO/TEI standard format that then constitutes the basis for all further processing steps. An encoder/decoder pair to/from WebLicht's TCF format allows existing services in WebLicht to be used, whereas additional services for spoken language data can operate directly on the ISO/TEI data. Section 2 sketches the salient characteristics of the ISO/TEI format. Sections 3 to 5 describe the different web services used in this workflow.

2 A TEI-Based ISO Standard for Spoken Language Transcriptions

The TEI Guidelines have always contained suggestions for representing "Transcriptions of Speech" (chapter 8). However, this portion of the guidelines has never been sufficiently established in the respective research communities to work as an interchange format or even a standard. One reason is that work with spoken language data relies on specialized tools for efficient transcription, and the relation of TEI to these tools was never sufficiently clarified. Schmidt (2005) made a first suggestion on how to reconcile the time-based data models, which most tools are based on, with a hierarchy-based data model that underlies the TEI proposal. Taking into account a tool interoperability study (Schmidt et al. 2009), the proposal for representing spoken language transcriptions on the basis of TEI was further refined in Schmidt (2011), and from 2012 on, became an ISO project (ISO/ TC 37/SC 4/WG 6), concluded in summer 2016 with the official publication of "Language resource management - Transcription of Spoken Language".

The focus of the standard is on orthographic (i.e. not: phonetic) transcription of recordings of authentic interaction (i.e. not: prompted speech or experiment data). Guiding design principles were the maxim to reuse as many elements as possible from the existing guidelines and orient their use towards interoperability with established tools. In particular, this meant restricting the choices wherever the guidelines offer more than one concept for representing a phenomenon.

```
<annotationBlock who="MJ" start="#T0" end="#T2" xml:id="ab1">
  <u xml:id="u1">
          <seg type="intonation-phrase" subtype="falling" xml:id="seg1">
                    <w xml:id="w1">I</w>
                    <vocal xml:id="voc1"><desc>cough</desc></vocal>
                    <w xml:id="w2">see/w>
                    <w xml:id="w3">a/w>
                    <w xml:id="w4">door/w>
          </seg>
  </u>
  <spanGrp type="lemma">
       <span from="w1" to="#w1">I</span>
       <span from="w2" to="#w2">see</span>
       <span from="w3" to="#w3">a</span>
       <span from="w4" to="#w4">door</span>
  </spanGrp>
  <spanGrp type="pos">
       <span from="w1" to="#w1">PPER</span>
       <span from="w2" to="#w2">V</span>
       <span from="w3" to="#w3">DET</span>
       <span from="w4" to="#w4">NN</span>
  </spanGrp>
</annotationBlock>
```

Figure 2. <annotationBlock> grouping an <u> with standoff annotation in <spanGrp>s

The central building blocks of the document are (a) a <timeline> defining offsets into a recording, (b) a <particDesc> defining the participants of the interaction and (c) a sequence of <u> elements, which correspond to individual speaker contributions and contain the actual transcription text as well as references to the timeline and to a participant. The standard allows for different levels of detail for the markup of the transcription text. In the simplest case, a plain text string can be used, which is temporally aligned via mandatory @start and @end attributes of the <u> element. Intervening temporal alignment can be added through <anchor/> elements. The microstructure of the speaker contribution can be represented by additional elements, most importantly <w> for word tokens, <pause> for pauses and <vocal> or

<kinesic> for non-verbal phenomena. Segmentations of speaker contributions into units above the word level can be represented by intervening <seg> elements. The additional markup below <u> is crucial for many automatic annotation methods.

For annotations on the primary transcription, standoff annotation layers in <spanGrp> elements can be used and grouped with the <u> element they belong to, using an <annotationBlock> element (Banski et al. 2016). This mechanism is crucial also for storing annotations that result from automatic methods. Figure 2 illustrates the annotation of an utterance with lemmas and POS tags. The standard was published by ISO in August 2016[1].

3 Converting Common Transcription Formats to ISO/TEI

Unlike other text types (such as manuscripts, dictionaries etc.) addressed by the TEI, spoken language transcription is rarely done by editing an XML document directly. Researchers crucially rely on tools that support the alignment of audio/video and transcription in an ergonomic graphical user interface. Hinrichs and Vogel (2010) identify ANVIL, CLAN, ELAN, EXMARaLDA, FOLKER, Praat and Transcriber as annotation tools that are currently relevant to the CLARIN community for this task. Most of them (CLAN and Praat being the exception) work with XML based formats, but, so far, none of them natively operates on a TEI compliant format.

In order to make the ISO/TEI standard work in practice, it is therefore essential to provide users with an easy way of converting from a tool format to the ISO/TEI format. The closer that tool format is to TEI's general structure, the more straightforward this conversion is. So far, we have developed converters for CLAN, ELAN, EXMARaLDA, FOLKER and Transcriber, including options to tokenise data according different transcription conventions such as cGAT, HIAT, CLAN. We are making the converters available as individual webservices, hosted by the CLARIN Centre at the Hamburg Centre for Language Corpora and made known via a CMDI description (e.g. PID 11022/0000-0000-9ABA-1 for the EXMARaLDA converter) to CLARIN.

4 Using TCF-based Web Services in WebLicht

(1)
```
<u who="MJ" start="#T0" end="#T2">
  <seg type="intonation-phrase" subtype="level">
    <w xml:id="w1">I</w>
    <vocal xml:id="voc1"><desc>cough</desc></vocal>
    <w xml:id="w2">see/w>
    <w xml:id="w3">a/w>
    <w xml:id="w4">door/w>
  </seg>
  <anchor synch="#T1"/>
  <seg type="intonation-phrase" subtype="falling">
    <w xml:id="w5">I/w>
    <pause xml:id="p1" dur="PT0.3S"/>
    <w xml:id="w6">want/w>
    <w xml:id="w7">to/w>
    <w xml:id="w8">paint/w>
    <w xml:id="w9">it/w>
    <w xml:id="w10">black/w>
  </seg>
</u>
```

(2)
```
<TextCorpus>
  <text>I see a door I want to paint it black</text>
    <tokens>
      <token ID="w1">I</token>
      <token ID="w2">see</token>
      <token ID="w3">a</token>
      <token ID="w4">door</token>
      <token ID="w5">I</token>
       <!-- [...] -->
       <token ID="w9">it</token>
       <token ID="w10">black</token>
    </tokens>
    <sentences>
       <sentence ID="s_1" tokenIDs="w1 w2 w3 w4"/>
       <sentence ID="s_2" tokenIDs="w5 w6 w7 w8 w9 w10"/>
    </sentences>
    <textSource type="application/tei+xml;
        format-variant=tei-iso-spoken;tokenized=1">
        <![CDATA[<TEI xmlns="http://www.tei-c.org/ns/1.0">
          [...]<u who="MJ" start="#T0" end="#T2">[...]</TEI>]]>
    </textSource>
<TextCorpus>
```

(3)
```
<TextCorpus>
    <!-- [...] -->
    <POStags tagset="stts">
          <tag ID="pt_0" tokenIDs="w1">PPER</tag>
          <tag ID="pt_1" tokenIDs="w2">V</tag>
          <tag ID="pt_2" tokenIDs="w3">DET</tag>
          <tag ID="pt_3" tokenIDs="w4">NN</tag>
          <!-- [...] -->
          <tag ID="pt_10" tokenIDs="w10">ADJ</tag>
        </POStags>
    <!-- [...] -->
</TextCorpus>
```

(4)
```
<annotationBlock who="MJ" start="#T0" end="#T2" xml:id="ab1">
  <u>
    <seg type="intonation-phrase" subtype="level">
      <w xml:id="w1">I</w>
      <vocal xml:id="voc1"><desc>cough</desc></vocal>
      <w xml:id="w2">see/w>
      <w xml:id="w3">a/w>
      <w xml:id="w4">door/w>
    </seg>
  </u>
  <spanGrp type="pos">
      <span from="#w1" to="#w1">PPER</span>
      <span from="#w2" to="#w2">V</span>
      <span from="#w3" to="#w3">DET</span>
      <span from="#w4" to="#w4">NN</span>
  </spanGrp>
</annotationBlock>
```

Figure 3. TCF encoding and decoding

[1] http://www.iso.org/iso/catalogue_detail.htm?csnumber=37338

Currently, all services integrated in WebLicht operate on the Text Corpus Format (TCF). In order for an ISO/TEI document to be sent through WebLicht, it has to be encoded in TCF before the chain is applied, and the result needs to be decoded from TCF to ISO/TEI. TCF is based on the "stream of tokens" idea: it assumes that the basic structure of any document is a linear sequence of token elements and "the tokens layer is [thus] the main anchor layer among TextCorpus layers, i.e. all other layers […] directly or indirectly […] reference tokens by referencing token identifiers." TCF thus does not have in its basic structure any means of representing some information that is crucial to spoken language transcription, such as time alignment and speaker assignment, and it also does not provide the possibility to distinguish different types of tokens (such as words vs. non-speech) on its basic layer. ISO/TEI-to-TCF conversion is therefore bound to be lossy – not all information in the source can be mapped onto some TCF element. The general approach for the TCF encoding of ISO/TEI files is therefore:

(1) To map only those elements of an ISO/TEI file that have a straightforward equivalent in TCF. Basically, this boils down to mapping <w> elements in ISO/TEI to <token> elements in TCF.
(2) To assume that most services in WebLicht will work fine with this reduced set of information.
(3) To keep on the <tokens> layer the @xml:id attributes of the original document.
(4) To keep the entire original ISO/TEI document in the <textSource> element.

When such a TCF document is fed through WebLicht, any resulting new annotation layers can be remapped to a suitable ISO/TEI form via the memorized id attributes. Since the original document in the <textSource> element remains unchanged, the decoding step can be stateless as required by CLARIN's SOA architecture. Figure 3 illustrates how an excerpt of an ISO/TEI transcription (1) is encoded in TCF (2), supplemented with the result of a POS tagger (3) and decoded back to ISO/TEI (4). The TCF encoder and decoder are made available as CLARIN compliant web services via the Hamburg Centre for language Corpora.

5 ISO/TEI-based Web Services

While TCF encoding and decoding is a workable solution for making existing services in WebLicht usable for spoken language data, it is not the optimal way of dealing with such data. The information lost in the TFC encoding process has potential value for automatic processing methods. For example, a POS tagger for spoken language might make use of n-gram statistics involving information about pauses, or a lemmatizer may treat in a special manner defect tokens like aborted words. In some cases, basic assumptions tacitly included in TCF may turn out to be overly simplified when applied to spoken language. For instance, TCF specifies "the language of the data" inside a single attribute @lang on the <textCorpus> element. Multilingual interactions, such as interpreted talk, however, contain by definition data in at least two languages, and there is no way of telling a TCF based tool which part of the document is in which language.

Another typical feature of transcriptions is the traditional use of various deviations from standard orthography to render characteristics of speech such as regular elisions and assimilations, specific varietal features, or idiosyncratic pronunciation. With a data model complex enough to handle both standard and non-stand renderings of single tokens, a standard orthographic rendering can be added and used as the basis for various automatic annotation tools. There are established workflows for this kind of normalization of transcriptions in use at the Archive for Spoken German at the IDS.

Ideally, automatic annotation tools and services optimised for spoken language transcription would therefore operate directly on the ISO/TEI format, and the detour via TCF conversion become unnecessary. An existing example for this is an adaptation of TreeTagger and the STTS for use with transcriptions from the FOLK project (Westpfahl/Schmidt 2016). We are currently working on turning this mechanism into a CLARIN compliant web service based on ISO/TEI. Similar tools developed in the context of corpora at the Archive for Spoken German at the IDS, such as the tool used for orthographic normalisation of transcriptions, and a tool for fine-aligning transcriptions via MAUS, could be treated analogously.

6 Concluding Remarks

As the present contribution shows, there are ways of making existing CLARIN components and architectural concepts that were originally or mainly developed for canonical written language data usable also for spoken language data. The issue of standardisation is crucial in this task, since, without a widely-used and sufficiently specified common basis, the number of processing steps needed to convert a given piece of data from and to a form usable by WebLicht would multiply. The newly published ISO/TEI standard provides such a common basis for various established tool formats and transcription conventions for spoken language data.

References

[Banski et al. 2016] Banski, P., Gaiffe, B., Lopez, P. Meoni, S., Romary, L., Schmidt, T., Stadler, P., Witt, A. (2016): *Wake up, standOff!*. Paper at the TEI Conference and Members' Meeting 2016, 26th to 30th September, Vienna, Austria.

[Hinrichs/Vogel 2010] Hinrichs, E., Vogel, I. (2010): *CLARIN - Interoperability and Standards.* CLARIN devliverable D5.C-3. http://www-sk.let.uu.nl/u/d5c-3.pdf

[Hinrichs et al. 2010] Hinrichs, M., Zastrow, T., Hinrichs, E. (2010): *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure.* In: Proceedings of LREC'10. Paris: ELRA.

[Kisler et al. 2010] Kisler, T. and Schiel, F. and Sloetjes, H. (2012): *Signal processing via web services: the use case WebMAUS*. In: Proceedings of Digital Humanities 2012, Hamburg, pp. 30-34.

[Menke et al. 2015] Menke P., Freigang F., Kronenberg T., Klett S., Bergmann K. (2015): *First Steps towards a Tool Chain for Automatic Processing of Multimodal Corpora.* Journal of Multimodal Communication Studies. 2:30-43.

[Schmidt 2005] Schmidt, T. (2005). *Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech.* Arbeiten zur Mehrsprachigkeit, Folge B, 62.

[Schmidt et al. 2009] Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Magnusson, M., Rose, T., Sloetjes, H. (2009). *An Exchange Format for Multimodal Annotations.* In Martin, J.-C., Paggio, P., Kipp, M., Heylen, D. eds., Multimodal Corpora (pp. 207-221). Springer.

[Schmidt 2011] Schmidt, T. (2011). *A TEI-based Approach to Standardising Spoken Language Transcription.* Journal of the Text Encoding Initiative, 1, 1-22.

[Schmidt/Wörner 2014] Schmidt, T., Wörner K. (2014): *EXMARaLDA.* In: Durand, J., Gut, U. and Kristoffersen, G. (eds.): The Oxford Handbook of Corpus Phonology. Oxford: OUP 2014, pp. 402-419.

[Westpfahl/Schmidt 2016] Westpfahl, S. / Schmidt, T. (2016): *FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German.* In: Proceedings of LREC'16. Paris: ELRA.