Originally published in: Simov, Kiril/Eskevich, Maria (Eds.): Proceedings of CLARIN Annual Conference 2019, 30 September - 2 October 2019, Leipzig, Germany. - Utrecht: CLARIN, 2019. Pp. 36-39.

CLARIN Web Services for TEI-annotated Transcripts of Spoken Language

Bernhard Fisseni, Thomas Schmidt
Leibniz-Institut für Deutsche Sprache (IDS)
Mannheim, Germany
{fisseni,thomas.schmidt}@ids-mannheim.de

Abstract

We present web services implementing a workflow for transcripts of spoken language following TEI guidelines, in particular ISO 24624:2016 "Language resource management – Transcription of spoken language". The web services are available at our website and will be available via the CLARIN infrastructure, including the Virtual Language Observatory and WebLicht.

1 Introduction / Recapitulation

Schmidt et al. (2017) sketch, and partly implement, an architecture for making CLARIN webservices usable for transcriptions of spoken language, focusing on the TEI-based standard ISO 24624:2016 "Language resource management – Transcription of spoken language" [henceforth **ISO/TEI**] as a pivot format for web services. The 2017 paper concentrates on a solution with an encoder/decoder pair which first transforms ISO/TEI to WebLicht's TCF[1] and re-transforms the TCF result of the service chain to ISO/TEI. Thus a large class of language technology tools becomes accessible to researchers working with spoken language while maintaining interoperability with tools which are commonly used for manual transcription and annotation of audiovisual material (such as ELAN, Praat or EXMARaLDA).

As Schmidt et al. (2017) argue, CLARIN's service-oriented approach could be further leveraged for spoken language data through the development of services which (a) take into account the specific characteristics of spoken data as well as the specific tasks arising in their curation, and which (b) operate directly on the ISO format without a 'detour' via TCF. The present contribution explores this option further with respect to a typical use case: curation of interview data (sec. 3), sketching a workflow for related use cases and describing a CLARIN-conformant implementation of this workflow (sec. 4).

2 Related Work

Workflows for the curation of interview data have been discussed in the CLARIN context on the occasion of several workshops on Oral History (https://oralhistory.eu/) where the focus of this work was on the use of speech technology (e.g. ASR, forced alignment) rather than on enriching textual transcription data with language technology, as in the current paper. Ideally, both approaches complement each other.

Several methods described here were originally developed in the context of the EXMARaLDA system (http://www.exmaralda.org), as part of the workflow for compiling the Research and Teaching Corpus of Spoken German (FOLK, see Schmidt 2016) and/or as components of curation workflows at the CLARIN B-centres Hamburg Center for Language Corpora (HZSK, https://corpora.uni-hamburg.de/) and the Archive for Spoken German at IDS (AGD, http://agd.ids-mannheim.de/). Details on the development of the POS tagging model are described by Westpfahl (2019). Several of the services described in sections 4 reuse and extend these methods (at least conceptually) and put them on a different technological basis thereby integrating them more fully into the CLARIN infrastructure.

Besides Schmidt et al. (2017) and the ISO specification itself, the role of TEI as a suitable basis of a standard for spoken language transcription has been discussed, among others, by Schmidt (2011) and Liégeois et al. (2017). The TEI guidelines' chapter 8 on "Transcriptions of Speech" has also been used in CLARIN resources such as the GOS Corpus of Spoken Slovene (see Verdonik et al. 2013) and as the basis for a CLARIN-wide format for parliamentary data.

[1] see https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

3 Use case: Legacy interview corpora

AGD hosts more than 80 spoken language corpora with more than 10,000 hours of audio or video recordings, increasing continuously through external collaborations and data deposits. AGD curates such external resources, i.e. putting audio/video recordings, metadata, transcripts and annotations into a state where they can be archived, discovered (= found) and reused. The main datatypes in AGD are *interaction corpora* (e.g. the FOLK corpus, Schmidt 2016), *variation corpora* (e.g. *Deutsch Heute* or *Australiendeutsch*), and *interview corpora*, on which we would like to focus in the present paper. An example of a curated interview corpus at AGD is Norbert Dittmar's *Berliner Wendekorpus*[2]. Currently under curation are, e.g., the audio recordings from an interview study on German refugees in Britain ("Kindertransporte", see Thüne 2019). Multilingualism plays a central role for these data because speakers have migration histories and recordings hence often include code switching or mixing. These data share a high potential for interdisciplinary reuse, e.g. in sociological or oral history studies, and similar data is curated in many other centres (mostly outside CLARIN).

Typical initial data deposits consist of audio, transcripts in modified orthography (in English, e.g., "dunno" for "don't know") in some word processor format, and more or less structured metadata in legacy formats. AGD curates such data (a) fully digitising the resource, (b) transforming textual data into structured, interoperable formats conforming to best practices and standards, (c) interconnecting different data types (e.g. aligning transcripts with recordings), (d) enriching data with linguistic information (e.g. POS-tagging), and (e) integrating them into DGD and into the wider language resource infrastructure (e.g. assigning PIDs, offering OAI/PMH). Curation workflows have common building blocks, which we propose to implement in a set of ISO/TEI-based, CLARIN-conformant web services. The same methods and tools can be used in a much wider range of contexts than just the specific use case illustrated here.

We use the following example[3] from the *Corpus Australian German* throughout the text, and show excerpts from the results of steps in the toolchain.

```
MC: Welche Früchte ham sie (.) hier in der (-) Gegend?
AS: Äh, Apfel. [...]
MC: Und ähm vielleicht könnten wir n bisschen umschalten ins Englische.
    What part of Germany did your forefathers come from?
AS: Eh, our people came from what they call Schlesien.
    I wouldn't know how you pronounce that in English.
```

4 Workflow and Tools

We provide an abstract description of the functionality of the services and an explanation of the motivation and challenges for each step.[4] The process is conceived of as a pipeline, so that the output of one step can immediately serve as input to the next step. We will also mention some parameters, but have to refer the reader to the documentation for a detailed description. All services can be given a default language which will be used if there isn't a more specific language annotation in the documents, or the language cannot be detected. Contrary to the approach in TCF, ISO/TEI documents thus inherently support multilingual texts.

4.1 Plain text to ISO/TEI-annotated texts (`text2iso`)

As detailed above, our use case regarding legacy corpora starts with documents in word processor format. We can disregard most of the formatting and expect input in plain text format for our web services. Hence the first step is to convert plain text transcribed data to a ISO/TEI-conformant format, which serves as input for all further processing steps.

In this step, the main challenge is to specify a plain text input format that is sufficiently expressive to serve in the most common cases, and sufficiently simple and restricted to be typed and parsed efficiently; parsing errors and other difficulties are signalled in XML comments. The format is supposed to allow segmentation of the conversation into utterances, and assignment of these utterances to speakers. A specification is available at `https://github.com/Exmaralda-Org/teispeechtools/blob/master/doc/Simple-EXMARaLDA.md`. The result of this step is a transcription file which is split into utterances: an `<annotationBlock>` for each utterance contains a `<u>` element as well as `<incident>` elements containing non-verbal actions and `<spanGrp>` elements containing commentaries. A `<timeline>` is derived from the text, and all annotation is situated

[2] `http://hdl.handle.net/10932/00-0332-BD7C-3EF5-0B01-4`, `http://agd.ids-mannheim.de/BW--_extern.shtml`

[3] `http://hdl.handle.net/10932/00-0332-BCFF-D7B3-7A01-9`, AD--_E_00010

[4] The web services are available at `http://clarin.ids-mannheim.de/teilicht`. The functionality is also available as a Java library and command-line tool, see `https://github.com/Exmaralda-Org/teispeechtools/`.

with respect to the <timeline>. Elements of the timeline are the beginning and end of each utterance; in case of overlap, the overlap start and end is referenced as an <anchor> within the utterances.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>
      <profileDesc><particDesc>
         <person n="AS" xml:id="AS"><persName><abbr>AS</abbr></persName></person>
         <person n="MC" xml:id="MC"><persName><abbr>MC</abbr></persName></person>
      </particDesc></profileDesc> <encodingDesc>...</encodingDesc> <revisionDesc>...</revisionDesc>
   </teiHeader> <text xml:lang="de">
      <timeline unit="ORDER">
         <when xml:id="B_1"/> <when xml:id="E_1"/>    <when xml:id="B_2"/> <when xml:id="E_2"/>
         <when xml:id="B_3"/> <when xml:id="E_3"/>    <when xml:id="B_4"/> <when xml:id="E_4"/>
         <when xml:id="B_5"/> <when xml:id="E_5"/>    <when xml:id="B_6"/> <when xml:id="E_6"/>
         <when xml:id="B_7"/> <when xml:id="E_7"/>    <when xml:id="B_8"/> <when xml:id="E_8"/>
      </timeline>
      <body>
         <annotationBlock start="B_1" end="E_1" who="MC">
            <u>Welche Früchte ham sie (.) hier in der (..) Gegend?</u>
         </annotationBlock> <annotationBlock start="B_2" end="E_2" start="B_2" who="AS">
            <u>Äh, Apfel.</u>
         </annotationBlock> ... <annotationBlock start="B_8" end="E_8" who="AS">
            <u>I wouldn't know how you pronounce that in English.</u>
         </annotationBlock>
      </body>
   </text>
</TEI>
```

4.2 Segmentation according to transcription convention (`segmentize`)

In the next step, the text is segmented according to transcription conventions. We enforce a tokenisation into words in <w> elements and punctuation in <pc>, and some information is lifted from the plain text of an <u> to the annotation level, mainly pauses and unclear or incomprehensible text. ISO/TEI allows to use time <anchor> elements also in the middle of words. Keeping the <anchor>s in place while processing the surrounding plain text was one of the challenges of implementing this step, as in this case, XML structure interferes with the abstract structure of the transcription.

```
<annotationBlock start="B_1" end="E_1" who="MC"><u>
  <w>Welche</w> <w>Früchte</w> <w>ham</w> <w>sie</w> <pause type="micro"/>
  <w>hier</w> <w>in</w> <w>der</w> <pause type="short"/> <w>Gegend</w> <pc>?</pc>
</u></annotationBlock>
```

4.3 Language detection (`guess`)

The motivation for this step is that interview data are often massively multilingual, and it is useful to be able to assign languages to single utterances. In contrast to TCF, the TEI formats allow `@xml:lang` on every structural level of text.

The service uses the Apache OpenNLP (`https://opennlp.apache.org/`) language models and language detector to process single utterances and guess what language they are in. It is possible to constrain the search space to a set of languages to avoid mis-detection of similar languages like German and Low German. A configurable threshold (default: 5 words) can be set to prevent language detection in utterances that are too short for a reliable result. In the result, the <u> have been annotated with `@xml:lang` attributes.

```
<annotationBlock start="B_5" end="E_5" who="MC"><u xml:lang="de">
     <w>Und</w> <w>ähm</w> <w>vielleicht</w> <w>könnten</w> <w>wir</w> ... </u>
</annotationBlock> <annotationBlock start="B_6" end="E_6" who="MC"><u xml:lang="en">
     <w>What</w> <w>part</w> <w>of</w> <w>Germany</w> <w>did</w> ... </u>
</annotationBlock>
```

4.4 OrthoNormal-like Normalisation (`normalize`)

EXMARaLDA includes the OrthoNormal tool for transcript normalisation, i.e. the mapping of tokens in modified orthography to their standard orthographic equivalent. The automated part of normalisation is dictionary-based and only available for German at the moment (see Schmidt 2012). Normalisation works on the <w> elements, which are annotated with a `@norm` attribute containing the normalised form.

```
<annotationBlock start="B_1" end="E_1" who="MC"><u xml:lang="de">
      <w norm="welche">Welche</w> <w norm="Früchte">Früchte</w> <w norm="haben">ham</w>
      <w norm="sie">sie</w> <pause type="micro"/> <w norm="hier">hier</w> <w norm="in">in</w>
      <w norm="der">der</w> <pause type="short"/> <w norm="Gegend">Gegend</w> <pc>?</pc>
</u></annotationBlock>
```

4.5 POS-Tagging with the TreeTagger (`pos`)

Preferrably after normalisation, a document can be part-of-speech-tagged and lemmatised. Tagging is done using TreeTagger via TT4J.[5] We use the standard tagging models provided by the TreeTagger, which are mainly for written language but include a model for spoken French, and additionally a model trained on spoken German by Westpfahl (2019). Respecting the language of the current word `<w>`, the correct parser model is chosen by language, and the `@pos` and `@lemma` attributes are set accordingly.

```
<annotationBlock start="B_5" end="E_5" who="MC"><u xml:lang="de">
  <w lemma="und" norm="und" pos="KON">Und</w> ... <w lemma="in" norm="ins" pos="APPRART">ins</w>
  <w lemma="Englische" norm="englische" pos="NN">Englische</w> <pc>.</pc>
</u></annotationBlock> <annotationBlock start="B_6" end="E_6" who="MC"><u xml:lang="en">
  <w lemma="what" pos="DTQ">What</w> ... <w lemma="come" pos="VVB">come</w> <w lemma="from" pos="PRP">from</w>
  <pc>?</pc>
</u></annotationBlock>
```

4.6 Pseudo-alignment using Phonetic Transcription or Orthographic Information (`align`)

Another addition to the EXMARaLDA workflow is pseudo-alignment between transcription and recordings using graphemic or phonemic information. Most of the data submitted to the paradigmatic workflow are not aligned, i.e. do not contain timestamps pointing from the transcript to the recording. A logical approach is to apply *forced alignment* on these. Several aligners exist, most importantly MAUS, provided by the Bavarian Archive for Speech Signals (BAS).[6] If possible, we use MAUS in our workflow. However, poor quality of the audio, large file sizes, or legal restriction can make this difficult or impossible. For such cases, a pseudo-alignment, which estimates an alignment based on the graph(em)ic form of utterances, i.e., counting letters or phone(me)s derived from a grapheme-to-phoneme conversion, is a useful alternative. Optionally, the canonical phonetic transcription can be added to the ISO/TEI document using the attribute `@phon` on `<w>` elements.

The alignment thus achieved can be manually improved, if necessary.

```
<timeline><tei:when interval="0s" xml:id="B_1"/> <tei:when xml:id="E_1" interval="5.394s" since="B_1"/>
  <tei:when xml:id="B_2" interval="5.394s" since="B_1"/> <tei:when xml:id="E_2" interval="6.356" since="B_1"/> ...
</timeline> <body>
  <annotationBlock end="E_2" start="B_2" who="AS"><u start="B_2" end="E_2">
    <w lemma="Äh" norm="äh" phon="ʔɛː" pos="ADJA">Äh</w> <pc>,</pc>
    <w lemma="Apfel" norm="Apfel" phon="ʔap.fəl" pos="NN">Apfel</w> <pc>.</pc>
  </u></annotationBlock> ...
```

5 Conclusion and Outlook

We hope to have shown that web services centred around ISO 24624:2016 form a useful addition to the CLARIN universe. The web services are currently available from IDS and will have been fully integrated into the CLARIN infrastructure by the time of the conference. Once the base architecture is thus established, we see several ways of evaluating and improving the individual services, e.g. by offering a direct choice of the tagger models for specific languages, or by testing whether language detection with moving windows can be applied to longer utterances in a way that detects language shifts like code switching.

References

Liégeois, L., Benzitoun, C., Etienne, C., & Parisse, C. (2017). Vers un format pivot commun pour la mutualisation, l'échange et l'analyse des corpus oraux. In *FLORAL*. Orléans, France.

Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. *Journal of the TEI*, *1*, 1–22.

Schmidt, T. (2012). EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In T. Declerck, K. Choukri, & N. Calzolari (Eds.), *Proceedings of LREC'12* (pp. 236–240). ELRA.

Schmidt, T. (2016). Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project. *Journal for language technology and computational linguistics*, *31*(1), 127–154.

Schmidt, T., Hedeland, H., & Jettka, D. (2017). Conversion and annotation web services for spoken language data in CLARIN. In L. Borin (Ed.), *Selected papers from the CLARIN annual conf. 2016* (pp. 113–130). Linköping University Electronic Press.

Thüne, E.-M. (2019). *Gerettet*. Berlin, Leipzig: Hentrich & Hentrich.

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation*, *47*(4), 1031–1048.

Westpfahl, S. (2019). *Dissertation (unpublished)* (PhD thesis). Universität Mannheim.

[5] see `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/` and `https://reckart.github.io/tt4j/`, respectively.

[6] `https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface`