IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Bernhard Fisseni, Thomas Schmidt

CLARIN WEB SERVICES FOR TEI-ANNOTATED TRANSCRIPTS

of Spoken Language

Leipzig, 2019-10-01, Parallel Session 2, 12:20

Mitglied der Leibniz-Gemeinschaft

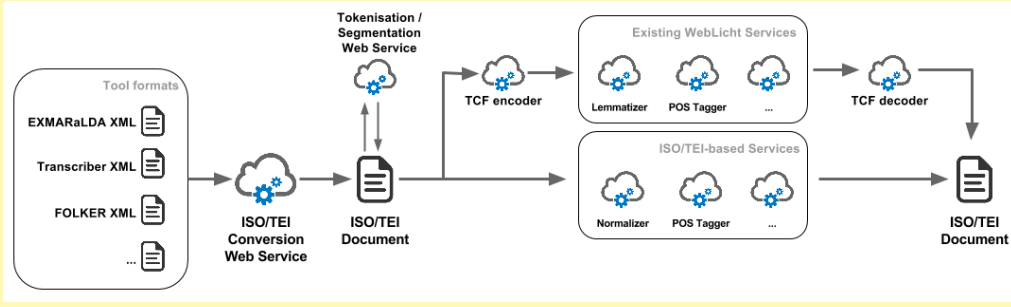
1 Introduction

Idea

CLARIN webservices operating on TEI for transcriptions of spoken language (Schmidt, Hedeland, and Jettka 2017)

pivot format TEI-based standard ISO 24624:2016 “Language resource management – Transcription of spoken language” [**ISO/TEI**]

approach 2017 WebLicht-oriented, detour to TCF



new perspective 2019 Avoid detour to TCF

use case interview corpora, with potential for extension

CLARIN-conformant implementation (practical problems notwithstanding)

Related Work

Workflows for the curation of interview data – workshops on Oral History (https://oralhistory.eu/)

focus on the use of speech technology (e.g. automatic speech recognition, forced alignment)

Workflow Elements from Other Contexts

- several methods originally developed in other contexts:
  - EXMARaLDA (http://www.exmaralda.org)
  - Research and Teaching Corpus of Spoken German (FOLK, see Schmidt 2016)
  - curation workflows at CLARIN B-centres HZSK (Hamburg) and Archive for Spoken German at IDS
  - POS tagging model described and developed by Westpfahl (2019)
- Here:
  - reuse and extend these methods
  - different technological basis
  - integrating more fully into the CLARIN infrastructure
  - library https://github.com/Exmaralda-Org/teispeechtools
  - web services https://github.com/Exmaralda-Org/teilicht

Related Work

TEI/ISO as Suitable Basis

- Schmidt (2011) and Liégeois, Benzitoun, Etienne, and Parisse (2017);
- GOS Corpus of Spoken Slovene (see Verdonik et al. 2013)
- a CLARIN-wide format for parliamentary data

2 Use case: Legacy Interview Corpora

Lecacy Interview Corpora I

Archiv für Gesprochenes Deutsch [Archive for Spoken German]

Curation, Presentation, Archival of corpora of spoken German

- > 80 spoken language corpora, > 10,000 hours of audio or video
- mainly
  - *interaction corpora* (e.g. the FOLK corpus, Schmidt 2016),
  - *variation corpora* (e.g. *Deutsch Heute* or *Australiendeutsch*),
  - ***interview corpora***,
    - Norbert Dittmar's *Berliner Wendekorpus*
    - Currently under curation: e.g., audio recordings from an interview study on German refugees in Britain ("Kindertransporte", see Thüne 2019).

Lecacy Interview Corpora II

Features of Legacy Interview Corpora

- typical initial data deposits:
  - audio
  - **transcripts in modified orthography** (in English, e.g., “dunno” for “don’t know”) in some word processor format
  - more or less structured **metadata** in legacy formats
- multilingualism ← code switching or mixing
- high potential for interdisciplinary reuse
- similar data in other centres (mostly outside CLARIN)

Curation

Common building blocks of curation workflows – partly also for other areas.

Typical Workflow

a. fully digitise the resource,
b. transform 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 the Database for Spoken German (DGD) and into the wider language resource infrastructure (e.g. assigning PIDs, offering OAI/PMH).

Example

[http://hdl.handle.net/10932/00-0332-BCFF-D7B3-7A01-9, AD–_E_00010]
from the *Corpus Australian German*

Plain Text Version

```
MC: Welche Früchte ham sie (.) hier in der (-) Gegend?
AS: Äh, Apfel.
    Apfel, Birnen, äh, Pflaumen, etwas Feigen, nich su viel und äh, dann hat \
    man auch äh Aprikosen, sehr viel Aprikosen und auch Pfirsiche, ja, \
    und äh, Mandeln sind auch sehr viel vorhanden.
    Mandeln tun eigentlich ganz gut hier.
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.
```

3 Workflows and Tools

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

Idea

- implementation: ANTLR 4 grammar
- challenge: plain text input format that is sufficiently expressive to serve in the most common cases of transcriptions
- `https://github.com/Exmaralda-Org/teispeechtools/blob/master/doc/Simple-EXMARaLDA.md`.

Input: plain text (Simple EXMARaLDA)

- speaker
- utterances
- accompanying actions etc., translations
- (pauses, unintelligible)

Parameters

- `lang`

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

Output: XML Structure

- comments with error messages
- one <annotationBlock> per utterance:
  - <u>
  - and <incident> elements containing non-verbal actions
  - <spanGrp> elements containing commentaries.
- a <timeline> is derived from the text
  - beginning and end of each utterance
  - simple overlaps, as <anchor> within the utterances

Plain Text Conversion Input

```
MC: Welche Früchte ham sie (.) hier in der (-) Gegend?
AS: Äh, Apfel.
    Apfel, Birnen, äh, Pflaumen, etwas Feigen, nich su viel und äh, dann hat \
    man auch äh Aprikosen, sehr viel Aprikosen und auch Pfirsiche, ja, \
    und äh, Mandeln sind auch sehr viel vorhanden.
    Mandeln tun eigentlich ganz gut hier.
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.
```

Plain Text Conversion Output XML: Header

```
<?xml version="1.0" encoding="UTF-8"?>
<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>
```

Plain Text Conversion Output XML: Timeline

```
<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>
```

Plain Text Conversion Output XML: Body

```
      <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" 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>
```

Segmentation according to transcription convention (`segmentize`)

Input

- TEI-conformant XML document
- containing `<u>` elements:
  - plain text according to a transcription convention (generic, cGAT minimal, cGAT basic)
  - potentially `<anchor>` elements referring to the `<timeline>`.
- **Challenge:** anchors in words

Parameters

- `lang` [local annotation will be preferred!]
- the transcription convention: `generic`, (cGAT) `minimal` and (cGAT) `basic`

Output

- TEI-conformant XML document
- the elements segmented into words
- conventions have been resolved to XML markup like `<pause>`, `<gap>` etc.

Segmentation Input and Output

Input

```
<annotationBlock start="B_1" end="E_1" who="MC">
  <u>Welche Früchte ham sie (.) hier in der (..) Gegend?</u>
</annotationBlock>
```

Output

```
<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>
```

Language detection (guess)

Idea

- interview data are often massively multilingual
- use Apache OpenNLP (`https://opennlp.apache.org/`) to do language detection

Input

- TEI-conformant with `<u>` and `<w>`

Parameters

- `expected languages`
- `lang` (fallback)
- threshold for `minimal` utterance length
- `force`

Language detection Output

```
<annotationBlock start="B_5" end="E_5" who="MC">
  <!--de: 0,07; en: 0,01; tr: 0,01--><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">
  <!--en: 0,05; de: 0,01; tr: 0,01--><u xml:lang="en">
      <w>What</w> <w>part</w> <w>of</w> <w>Germany</w> <w>did</w> ...
  </u>
</annotationBlock>
```

OrthoNormal-like Normalisation (`normalize`)

Algorithm (Schmidt 2012) – currently works only for German

1. apply most frequent normalisation for a word form in the FOLK corpus
2. OR consult list of words that occur capitalized-only in DeReKo (*Deutsches Referenzkorpus*)
3. (Out-of-dictionary words left as is)

Parameters

- `lang` (fallback)
- `force`

Input

- TEI-conformant XML with `<w>`

Output adds

- `@norm` attribute

Normalisation Output

```
<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>
```

POS-Tagging with the TreeTagger (pos)

Idea

Use TreeTagger [Helmut Schmid] and TT4J [Richard Eckart de Castilho]
http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ and https://reckart.github.io/tt4j/
e.g., model for spoken French from TreeTagger, model trained on spoken German by Westpfahl (2019)

Input

- TEI-conformant with <w>

Parameters

- lang (fallback)
- force

Output adds

- @lemma
- @pos

POS-Tagging Output

```
<annotationBlock start="B_5" end="E_5" who="MC"><u xml:lang="de">
  <w norm="und" lemma="und" pos="KON">Und</w>
  ...
  <w norm="ins" lemma="in" pos="APPRART">ins</w>
  <w norm="englische" lemma="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>
```

Pseudo-alignment (`align`) I

Idea

- forced alignment not always possible:
  - insufficient audio
  - data sensitivity
  - problems with large files (> 10 minutes?)
- use length of **phonetic transcription** or **orthographic representation** as an estimate of utterance length

Pseudo-alignment (`align`) II

Parameters

- `lang` (fallback)
- `force`
- `time`: overall length
- `use_phones` or graphs
- `transcribe`: add transcription?, only if `use_phones`
- `offset`: time of first event
- `every`: number of items after which to insert anchors

Pseudo-alignment (`align`) III

Output

- `<timeLine>` segmented according to estimates of orthographic strings
- `<timeLine>` segmented according to estimates of phonetic strings
  - counting phones, including length markers
  - disregarding syllable marks
  - falling back on orthography
  - respect lengths of `<pause>`s
- `@phon`

Transcription

- BAS' `runG2P` `http://clarin.phonetik.uni-muenchen.de/BASWebServices/help`
- Challenge: language tags, ISO-639-3 or ISO-639-2, very specific tags

Pseudo-alignment (`align`) IV

Pseudo-alignment Output

```
<timeline><tei:when interval="0s" xml:id="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> ...
```

4 Conclusion and Outlook

Conclusion and Outlook

Conclusion

- TEI-based, multilingual web services work as a proof of concept.
- They also have practical value for the curation of interview corpora.

Future Work

- Broader applicability to sub-classes of TEI documents?
  - e.g., now presuppose `<w>` and language annotation only for tagging
- Improve services:
  - other models, e.g. specific models for POS-tagging
  - e.g. moving windows for detecting language shifts
- More services?

https://clarin.ids-mannheim.de/teilicht

References

Liégeois, Loïc, Christophe Benzitoun, Carole Etienne, and Christophe Parisse (Mar. 2017). “Vers un format pivot commun pour la mutualisation, l’échange et l’analyse des corpus oraux”. *FLORAL*. Orléans, France.

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

– (2012). “EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language”. *Proceedings of LREC’12*. Ed. by Thierry Declerck, Khalid Choukri, and Nicoletta Calzolari. ELRA, pp. 236–240. ISBN: 978-2-9517408-7-7.

– (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. Ed. by Marc Kupietz and Alexander Geyken, pp. 127–154. ISSN: 2190-6858.

Schmidt, Thomas, Hanna Hedeland, and Daniel Jettka (2017). “Conversion and annotation web services for spoken language data in CLARIN”. *Selected papers from the CLARIN Annual Conf. 2016*. Ed. by Lars Borin. Linköping University Electronic Press, pp. 113–130.

Thüne, Eva-Maria (2019). *Gerettet*. Berlin, Leipzig: Hentrich & Hentrich.

Verdonik, Darinka, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko Stabej (Dec. 2013). “Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS”. *Language Resources and Evaluation* 47.4, pp. 1031–1048. ISSN: 1574-0218.

Westpfahl, Swantje (2019). “Dissertation (unpublished)”. PhD thesis. Universität Mannheim.