Conversion and Annotation Web Services for Spoken Language Data in CLARIN

Thomas Schmidt,
Archive for Spoken German,
Institute for the German Language, Mannheim

Hanna Hedeland & Daniel Jettka,
Hamburg Centre for Language Corpora,
University of Hamburg

Web Services in CLARIN(-D)

- Lemmatizing, POS Tagging, Named Entity Recognition, Parsing, …
- Annotation chains in WebLicht
- Built with, meant to operate on „canonical“ written language
- Text Corpus Format (TCF) as the underlying data model

➔ **Use with (transcriptions of) spoken language?**

Spoken language data

- A few widely used formats – CHAT, EAF, EXMARaLDA, Transcriber, Praat, … (see *CLARIN - Interoperability and Standards,* D5.C-3)
- More than / different from a “stream of tokens”
  - Non-Speech “tokens” (pauses, non-verbal descriptions, breathing)
  - Parallel structures (overlaps, alternative transcriptions)
  - Time alignment
  - No sentences, no document hierarchy
  - Defect tokens (incomprehensible speech, incomplete words, disfluencies)

➔ TCF not sufficient to accommodate all information

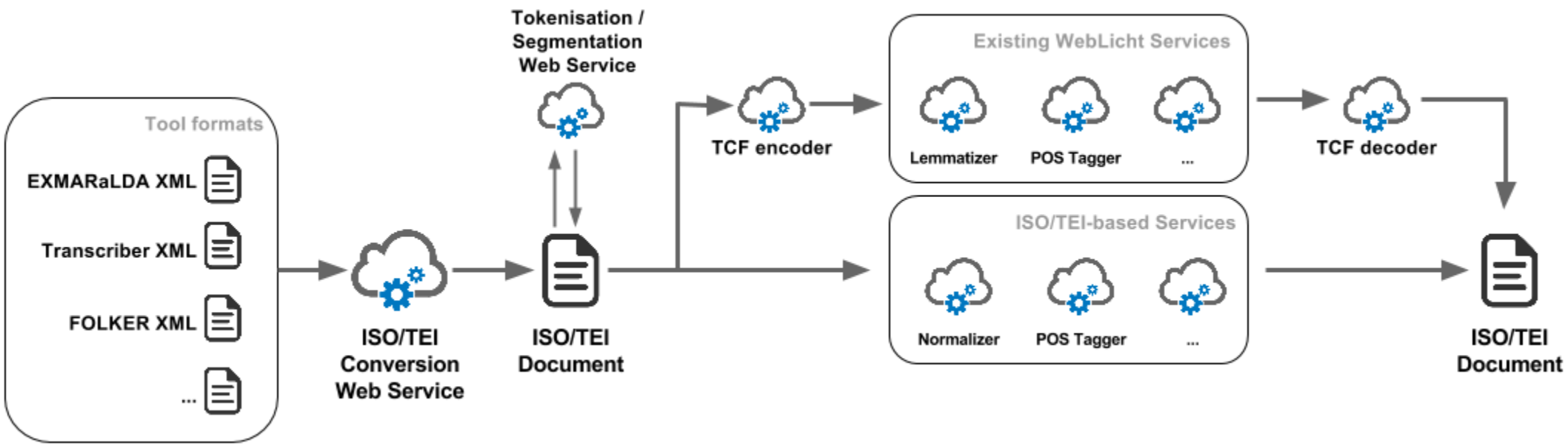
General approach

(1) A common format to represent transcriptions
- ISO/ TC 37/SC 4/WG 6: “Language resource management - Transcription of Spoken Language”, based on TEI guidelines
- Conversion web services from tool formats to ISO/TEI

(2a) Encode to TCF – use existing services – decode from TCF
- „Hide“ spoken language specifics
- (Tokenisation web service), Codec web service

(2b) Spoken language specifc services
- Orthographic normalisation, forced word aligment, prosodic annotation, …
- „pause-aware“ POS tagging

Architecture



ISO „Transcription of Spoken Language“

- Schmidt (2005), Schmidt et al. (2008), CLARIN guide on interoperability, Schmidt (2011): TEI for spoken language, tool interoperability
- ISO project 2012 to 2016, published in August 2016
- Scope: Orthographic transcription, verbal behaviour
- Ready to use
  - Compatible with data at HZSK and AGD
  - Existing converters for EXMARaLDA, FOLKER, Transcriber, CHAT
  - EAF conversion on corpus-for-corpus basis
  - Praat conversion with intermediate step

ISO „Transcription of Spoken Language“

Macro structure

- Speakers - <particDesc>
- Timeline - <timeline>, <when>
- Sequence of <u> with @start and @end
- <anchor> for arbitrary alignment

Micro structure

- Optional (→ CDATA only)
- „Tokenisation“ - <w>, <pause>, <vocal>, …
- Segmentation - <seg>

```
(1) <u who=”MJ” start=”#T0” end=”#T2”>
      I ((cough)) see a door. I (0.3) want to paint it (black/blue).</u>


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


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


(4) <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><w>black/w><w>blue</w></choice></unclear>
      </seg></u>
```

ISO „Transcription of Spoken Language“

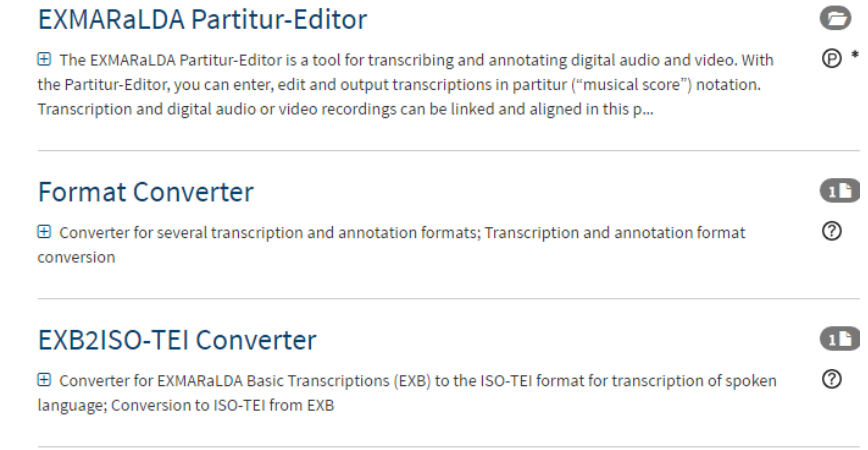
Standoff annotation

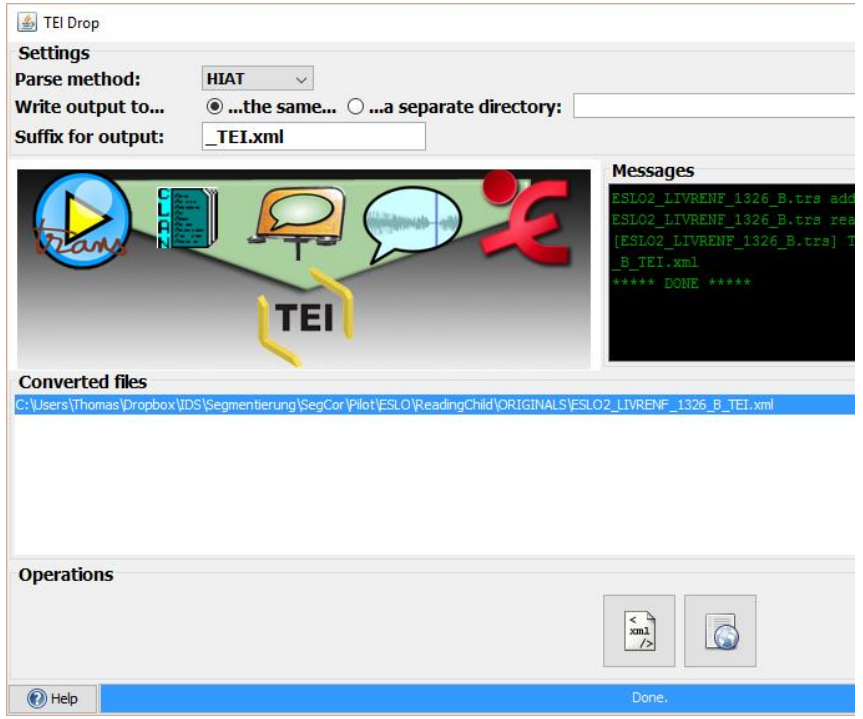
- Generic mechanism: <spanGrp>, + ID/IDREFS
- <annotationBlock> to group utterances with their annotations
- Borrowed from TEI standoff proposal (L. Romary, Banski et al. 2016)

```
<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>
```

Step 1: Tool format to ISO/TEI

- XSL stylesheets for EXMARaLDA, Transcriber, FOLKER
- Conversion „by proxy“ (through EXMARaLDA) for CHAT, Praat
- Import/Export filters in the tools
- TEI Drop as a desktop application („Droplet“)
- CLARIN Web services at HZSK (e.g. PID 11022/0000-0000-9ABA-1)
- N.B.: No explicit markup of micro structure in tool formats (except for FOLKER), but: Tokenisation obligatory for use with TCF ➔ „Segmentation“ algorithms for GAT, HIAT, CHAT conventions





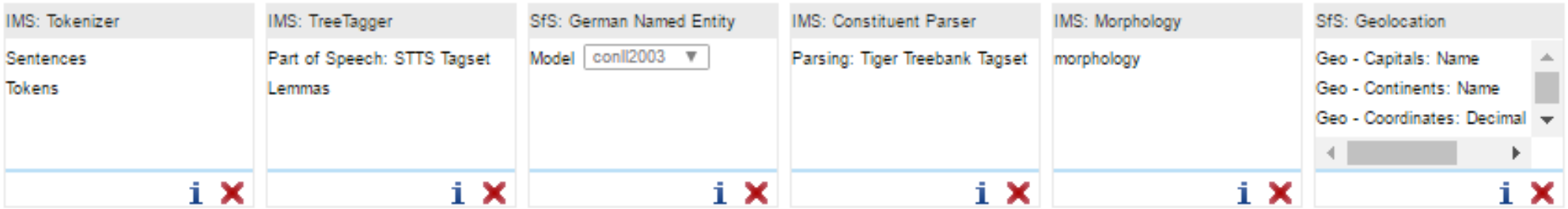
```
<u who="MJ" start="#T0" end="#T2">
   <seg type="intonation-phrase" subtype="level">
     <w xml:id="w1">I</w>
     <vocal xml:id="voc1"><desc>cough</desc></vocal>
     <w xml:id="w2">see/w>
     <w xml:id="w3">a/w>
     <w xml:id="w4">door/w>
   </seg>
</u>
```

```
<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>
    </tokens>
    <sentences>
       <sentence ID="s_1" tokenIDs="w1 w2 w3 w4"/>
    </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>
```

Step 2: ISO/TEI to TCF input (encoding)

- Map what can be mapped
  - basically <w> ➔ <token>
  - <u> or <seg> as sentence equivalents
- Keep original document in <textSource> (for stateless decoding)
- Keep original IDs (for inserting new annotations)

Step 3: WebLicht Chain…



```
<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>
```

```
<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>
```

Step 4: TCF output to TEI/ISO (decoding)

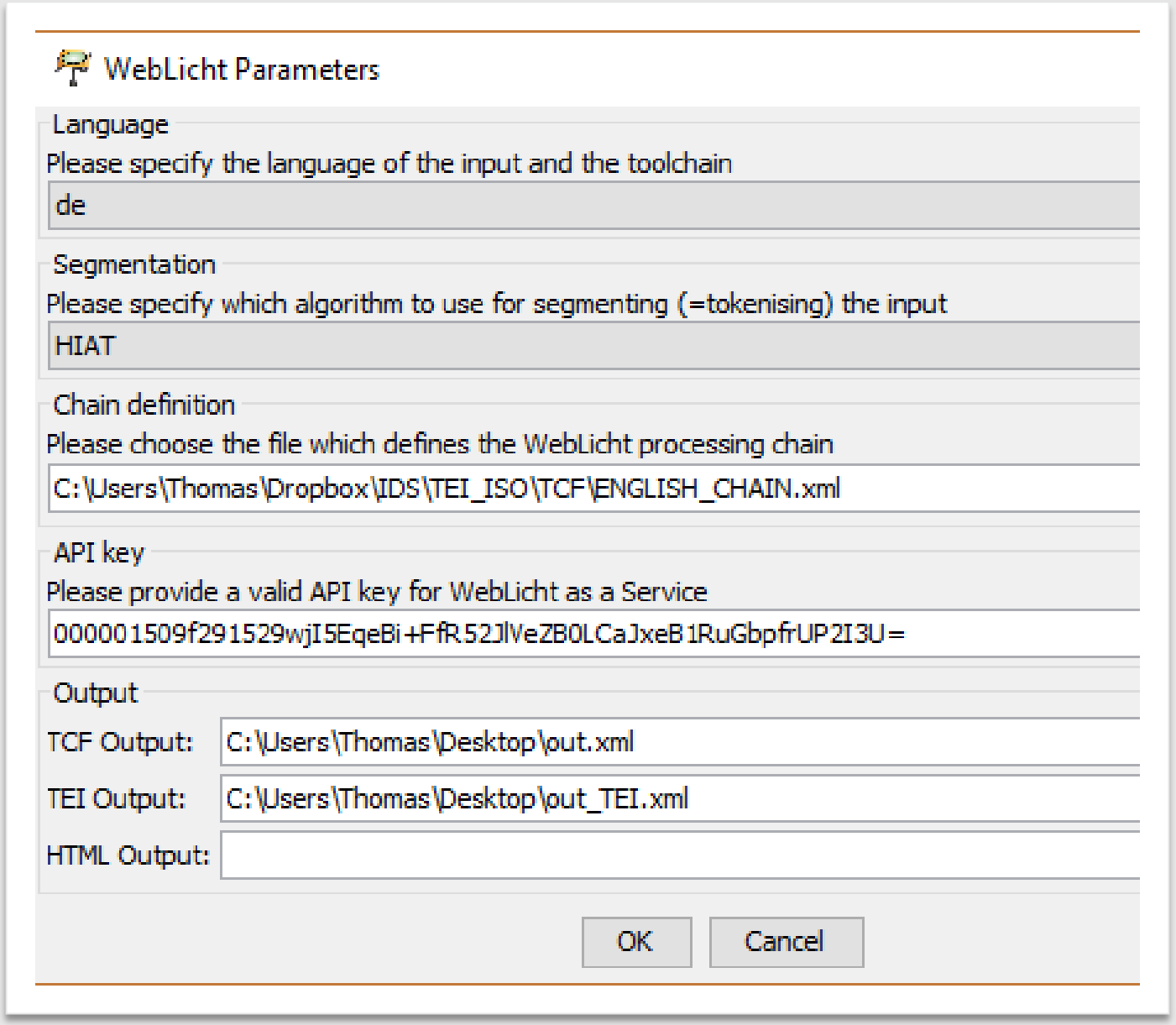
- Map TCF annotation layers to <spanGrp> /

Or: Cut out the middleman!

- No detour via TCF, no information loss
- Web services operating directly on TEI/ISO format
  - e.g. normalisation layer (modified orthography ➔ standard orthography, FOLK project)
  - e.g. POS tagging for interaction data (STTS 2.0 with TreeTagger, Westpfahl/Schmidt 2016)
  - also: services operating on the audio signal, e.g. forced word aligment through WebMAUS
- Currently no appropriate chaining tool (but Switchboard?)

Implementation status

- Proof of concept in EXMARaLDA (through WaaS)
- Converters ready
- WebLicht integration coordinated (MIME-types for format information!)
- Web services under construction



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.