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
   • 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 specific services
   • Orthographic normalisation, forced word alignment, prosodic annotation, ...
   • „pause-aware“ POS tagging
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>`, `<span>` + ID/IDREFS
- `<annotationBlock>` to group utterances with their annotations
- Borrowed from TEI standoff proposal (L. Romary, Banski et al. 2016)
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
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="abl">
  <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> / <span>
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 alignment 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


