CLARIN Web Services for TEI-annotated Transcripts of Spoken Language

Bernhard Fisseni
Leibniz-Institut für Deutsche Sprache (IDS)
Mannheim, Germany
fisseni@ids-mannheim.de

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

Abstract

We present web services which implement a workflow for transcripts of spoken language following the 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

Spoken language corpora are an important type of language resource as they exhibit many interesting aspects of language, such as language as a means of social interaction, dialectal, sociolectal or other forms of variation that are not covered by their written counterparts. Spoken language corpora require specialised processing methods. When dealing with audiovisual recordings, these methods can be based on speech technology; when dealing with the transcriptions of such recordings, methods can be used that are also applicable to written language data. The present paper focusses on the latter type of processing. It proposes elements of a workflow with documents in the TEI-based standard ISO 24624:2016 ”Language resource management – Transcription of spoken language” (henceforth **ISO/TEI**, see ISO 2016).

Schmidt, Hedeland and Jettka (2017) sketch, and partly implement, an architecture for making CLARIN webservices usable for transcriptions of spoken language, focusing on ISO/TEI as a pivot format on which web services operate and in which their output annotations can be represented. Schmidt, Hedeland and Jettka (2017) concentrates on a solution with an encoder/decoder pair which, at the entry point to a web service chain, transforms the ISO format to the Text Corpus Format TCF, which has been established as the basis for tool interoperability in WebLicht (see E. Hinrichs, M. Hinrichs and Zastrow 2010), and re-transforms the TCF result of the chain to ISO/TEI at the exit point. Since converters from common tool formats, such as those used by EXMARaLDA (see Schmidt and Wörner 2014), FOLKER (see Schmidt and Schütte 2010), ELAN (see Sloetjes 2014), Transcriber (see Barras et al. 2001) or CLAN/CHAT (see MacWhinney 2000), to ISO/TEI exist and can be prepended to the chain, a large class of language technology tools developed for written data thus 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.

Schmidt, Hedeland and Jettka (2017) argue in the outlook tat CLARIN’s service-oriented approach could be further leveraged for spoken language data through the development of adequate services. These must take into account the specific characteristics of transcribed spoken data. Important features are the use of forms deviating from standard orthography and the fact that multilinguality is a much more frequent phenomenon in spoken data (see the third paragraph of section 3). Moreover, these services must be adapted to the specific tasks arising in the curation of oral corpora, such as the alignment between transcript and audio. These services could operate directly on the ISO format, which provides features to cater for the aforementioned features of spoken corpora, without a ‘detour’ to TCF. The work reported in the present contribution explores this option further, first, by portraying a use case that typically arises in the curation of interview data (sec. 3), second, by sketching elements of a workflow suitable for this and related use cases and describing details of a CLARIN-conformant implementation of this workflow (sec. 4).

Bernhard Fisseni and Thomas Schmidt 2020. CLARIN Web Services for TEI-annotated Transcripts of Spoken Language. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 12–22.

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 data, whose results are documented on a dedicated website (`https://oralhistory.eu/`). The focus of this work is on the use of speech technology (e.g. automatic speech recognition, forced alignment) which operates on the audio signal. By contrast, the current paper concentrates on tools for enriching textual transcription data with language technology. Ideally these two approaches should be combined to complement each other.

Several methods described here were originally developed in the context of the EXMARaLDA system (Schmidt and Wörner 2014), 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-D B-centres Hamburg Center for Language Corpora (*Hamburger Zentrum für Sprachkorpora*, HZSK)[1] and the Archive for Spoken German at IDS (*Archiv für gesprochenes Deutsch*, AGD, Schmidt 2017).[2] Details on the development of the POS tagging model are described by Westpfahl (2020). 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, Hedeland and Jettka (2017) and the ISO specification itself (ISO 2016), 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" (TEI Consortium 2019) 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.[3]

3 Use case: Legacy interview corpora

The Archive for Spoken German (*Archiv für Gesprochenes Deutsch*, AGD) at the Leibniz Institute for the German Language is a central point for depositing, archiving, and disseminating corpora of spoken German. AGD hosts more than 80 spoken language corpora with more than 10,000 hours of audio or video recordings. The archive's stock is increasing continuously through internal corpus compilation projects, through collaborations with external partners and through data deposits by completed projects. A substantial part of the archive's work goes into curating 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 (thus conforming to the FAIR principles, cf. Wilkinson et al. 2016), among others through CLARIN services like the Virtual Language Observatory (VLO). The data types which the AGD deals with can be roughly divided into three classes:

(1) *interaction corpora* which document language in interaction (e.g. the FOLK corpus, Schmidt 2016),

(2) *variation corpora* which deal with language variation (dialects, regiolects, etc.) within the German-speaking core countries (e.g. *Deutsch Heute*, Brinckmann et al. 2008) and in German language communities around the world (e.g. Australian German, Lich and Clyne 1984) and

(3) *interview corpora*, which consist of relatively free (mostly narrative or biographic) interviews with selected speaker groups and/or on specific topics. In the present paper, we would like to focus on this corpus type.

Examples for already curated interview corpora at the AGD are Norbert Dittmar's *Berliner Wendekorpus* (see Schmidt 2019)[4] in which speakers from East and West Berlin were asked to relate their personal experiences with the fall of the Berlin wall, Anne Betten's extensive data on German-speaking emigrants to Israel (see Betten 1995),[5] or a recent interview study by Serap Devran (see Devran 2017) which deals with people of Turkish descent who (re)migrated to Turkey from Germany (available in the DGD since January 2020 since January 2020). It should be pointed out that ("language-biographic") interviews are also often an integral part of variation corpora. It is also worth noting that multilingualism plays a central role for a substantial part of these data because the respective studies focus on speakers with migration histories and their specific language varieties which often include code switching or mixing.

[1] see `https://corpora.uni-hamburg.de/`

[2] see `http://agd.ids-mannheim.de/`

[3] see `https://www.clarin.eu/event/2019/parlaformat-workshop`

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

[5] see `http://hdl.handle.net/10932/00-0332-C3A7-393A-8A01-3`, `http://agd.ids-mannheim.de/ISW-_extern.shtml`

While they cover a wide range of topics, these data have a lot in common in terms of methodology (all of them are semi-structured, narrative or biographic interviews with little interviewer intervention and a high degree of spontaneity) and in terms of structural and technical properties (typically audio recordings in quiet environments with durations up to three or four hours, rich biographic metadata on the speakers, orthographic transcriptions). They also share a high potential for interdisciplinary reuse, mostly because, beyond documenting specific linguistic forms, their contents also make them a valuable source for sociological or oral history studies.

The AGD has recently acquired, or is in the process of acquiring, further such interview corpora. Among them are the 2800 hours of audio recordings from the Bonn Longitudinal study of Aging (*Bonner gerontologische Längsschnittstudie*, see Lehr and Thomae 1987) and an interview study on German refugees in Britain ("Kindertransporte", see Thüne 2019). Other projects or data centres (outside of CLARIN) in Germany dealing with similar data are *Zwangsarbeit-Archiv* at the *Center für Digitale Systeme* (CeDis) in Berlin[6], *Archiv „Deutsches Gedächtnis"* at *FernUniversität in Hagen* (the German distance-learning university)[7] and *QualiService Bremen*[8], a centre for qualitative research data in the social sciences.

Typically, when data from interview studies are deposited at the AGD, they consist of the audio recordings (digitised or not), transcripts in modified orthography (e.g. "zwohunnert" for a spoken form of standard "zweihundert" ('two hundred') ; "dunno" for "don't know" would be a similar example in English) in some word processor format, and more or less structured metadata on interviews and interviewees in spreadsheet, text files, or similar formats. The AGD curates such data with the aims of:

(a) fully digitising the resource, especially the primary audio or video data,
(b) transforming all textual data into structured, machine-readable, interoperable formats which conform to current best practices and/or standards (e.g. for transcripts ISO/TEI; for metadata CMDI, cf. Broeder et al. 2012; CLARIN ERIC 2019),
(c) interconnecting the different data types (e.g. aligning transcripts with recordings, referencing between primary, secondary and metadata),
(d) enriching the data with further information useful for linguistic analysis (e.g. lemmatisation, POS tagging) and
(e) integrating them into the Database for Spoken German (DGD, `https://dgd.ids-mannheim.de`) as the primary dissemination platform and
(f) integrating them into the wider language resource infrastructure through the institute's long term archiving repository thereby associating datasets with PIDs and making metadata available through OAI/PMH.

The workflows needed for transforming the deposited source data into the desired target state may differ with the details of the individual resource, but experience has shown that they are always made up from the same set of building blocks. It is these building blocks that we propose to implement in a set of ISO/TEI-based, CLARIN-conformant web services and which we will describe in more detail in the following sections. While we illustrate them on a specific piece of data from a specific data centre, we argue that the same methods and tools, if properly configured, will be useful and applicable in a much wider range of contexts.

For example, the output of many transcription tools such as ELAN, EXMARaLDA, FOLKER, Transcriber, CLAN or Praat can be directly transformed to ISO/TEI with the help of suitable conversion tools (such as the web services described by Schmidt, Hedeland and Jettka 2017). Skipping the first step in the toolchain, viz. plain text to ISO/TEI conversion, the tools described here can be used on these data, thus applying to a much wider range of TEI documents than those deriving from our use case.

Furthermore, not all methods described here require the input to be fully conformant ISO/TEI transcripts. The language detection service, for example, can be applied to any data which uses the TEI <u> (*utterance*)[9] in a meaningful way as an annotation. Likewise, the only prerequisite for lemmatization and POS-tagging is that texts have been tokenized with TEI <w> (*word*)[10] elements.

[6] see `https://www.zwangsarbeit-archiv.de/index.html`

[7] see `https://deutsches-gedaechtnis.fernuni-hagen.de/`

[8] see `https://www.qualiservice.org/`

[9] see `https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-u.html`

[10] see `https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-w.html`

Normalisation can also be applied to any <w>-level annotated texts, but it is probably only useful in cases of playful writing, e.g. in computer-mediated communication. Even then, it would probably be preferable to make a new dictionary of normalisation and capitalised-only words.

We use the following example from the *Corpus Australian German* (see Clyne 1981; Kipp 2002)[11], gathered by the Australian linguist Michael Clyne in the 1960s and deposited at the AGD in 2014, throughout the text, and show excerpts from the results of step in the toolchain. The backslash indicates that a line is only broken for typesetting purposes.

```
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.
```

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.[12]

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 we 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 the language of the document cannot be otherwise determined. Contrary to the approach in TCF, ISO/TEI documents, and TEI documents in general, inherently support multilingual texts, that is: not only can a language be specified for the text as a whole, but individual components (here: utterances or words) can be assigned differing language tags.

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

As detailed in sec. 3, our use case of legacy corpora starts with documents in word processor format. As we can disregard most of the formatting, we 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.[13]

In this step, the main challenge was specifying a plain text input format that is sufficiently expressive to serve in the most common cases of transcriptions that will be subject to the workflow, as outlined above, and sufficiently simple and restricted to be typed and parsed efficiently. Conventions should also be as close as possible to those typically used in the text submitted to the AGD. The latter point was a reason to exclude existing formats such as CHAT. The format is supposed to allow segmentation of the conversation into utterances, assignment of these utterances to speakers. A specification is available at `https://github.com/Exmaralda-Org/teispeechtools/blob/master/doc/Simple-EXMARaLDA.md`.

Building on previous work, we spelt out some restrictions and corner cases and specified a formal language which can be deterministically parsed. Parsing was implemented using the ANTLR 4 parser generator.[14] The format manages simple forms of overlap between utterances as well as the annotation of nonverbal actions accompanying or stepping in for verbal actions. As we did not want to add too much explicit markup, we had to specify limitations with respect to overlap handling. As can be seen from the example, overlaps are indicated by marks occurring in the text. The restriction is that such marks can occur freely in the first utterance containing them, but to avoid complicated temporal alignment structures that might turn contradictory, marks must occur at the beginning of later utterances referencing them.

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

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

[12] The web services are available at `http://clarin.ids-mannheim.de/teilicht`.

[13] For intergration into WebLicht, see sec. 4.8, `text2iso` and `segmentize` were combined into a service `text2seg`, which takes all the parameters of these services.

[14] see `https://www.antlr.org/`

<spanGrp> elements containing commentaries. A <timeline> is derived from the text, and all annotation is situated 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. Again, this is implemented deterministically by processing a formal language. We enforce a a tokenisation into words in TEI <w> elements and punctuation in TEI <pc>, and some information is lifted from the plain text of an <u> to the annotation level, mainly pauses (encoded as TEI <pause> with a @type attribute) and unclear or incomprehensible text. The most adequate transcription level for the paradigmatic workflow is the *generic* transcription level, which provides these basic features.[15] More advanced transcription levels follow cGAT conventions.[16]

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)

Language detection is an addition to the workflow implemented in EXMARaLDA up to now. The motivation for this step is that, as mentioned in sec. 3, data are often massively multilingual, and it is useful

[15] The specification is available at https://github.com/Exmaralda-Org/teispeechtools/blob/master/doc/Generic-Conventions.md

[16] see http://agd.ids-mannheim.de/gat.shtml

to be able to assign languages to single utterances. In contrast to TCF, the TEI formats allow @xml:lang to specify a language on every structural level of text. We leverage this attribute to annotate language changes.

The service uses the Apache OpenNLP[17] 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; the default is German, Turkish and English. Language detection quality deteriorates if too little linguistic material is available, and if the transcription deviates too much from standard orthography. Therefore, a configurable threshold (default: 5 words) can be set to prevent potentially unreliable language detection in utterances that are too short.

In the result, the <u> have been annotated with @xml:lang attributes where the algorithm[18] reached a decision. If languages are equally probable, the document language is preferred. Cases of doubt are reported in XML comments; for debugging purposes, we also report probabilities for the expected languages here.

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

The following steps depend on correct language classification, and can hence be facilitated by manual language annotation or by applying guess before they are executed.

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, e.g. “zwohunnert” to “zweihundert”, “kannste” to “kannst Du”, “hab isch net” to “habe ich nicht”, but also nouns which are non-capitalized according to the transcription convention, but capitalized according to standard orthography (“haus” to “Haus”).

The automated part of normalisation is dictionary-based and only available for German at the moment (see Schmidt 2012). We plan to experiment with other algorithms or languages in the future.

Normalisation works on the <w> elements, which are annotated with a @norm attribute containing the normalised form. The algorithm can be summarised as follows:

1. The most frequent normalisation for a word form in the FOLK corpus is applied.
2. If nothing is found in Step 1, the list of words that occur capitalized-only in DeReKo[19] is consulted and a normalisation is chosen.
3. Out-of-dictionary words are left as is.

On the FOLK corpus, this automatic procedure yields correct normalisations for 93% of all tokens, and for 83% of tokens which require normalisation. Since the FOLK corpus is relatively large and contains data from diverse regions and settings, we can expect the procedure to perform with similar quality on interview data.

```
<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>
```

[17] see https://opennlp.apache.org/

[18] see https://opennlp.apache.org/docs/1.9.0/manual/opennlp.html#tools.langdetect

[19] *Deutsches Referenzkorpus*, see http://www1.ids-mannheim.de/kl/projekte/korpora.html

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

POS-tagging and lemmatisation are preferrably done after normalisation, since a notably higher precision is achieved when the tagger is fed normalised forms instead of forms in modified orthography (see Westpfahl 2020). However, it is not a requirement for this step that transcripts be normalised. We use the TreeTagger by Helmut Schmid (1995) for POS tagging, employing the Java wrapper TT4J by Richard Eckart de Castilho.[20]

We use the standard tagging models provided by the TreeTagger, which were mostly trained on and intended for written language. Tagging models trained on and intended for spoken language exist for French and for German (Westpfahl 2020). As Westpfahl (2020) shows for German, tagging models trained on spoken language data and with tag sets optimised for this resource type will yield significantly lower error rates (around 5% as compared to 15%–20%) than tagging models which have not been adapted to this task.[21]

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>
```

Note how in our example, this results in different tag sets being used for `<u>` elements in different languages.

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 do not contain information on the time when utterances occurred.

A logical step would be to apply *forced alignment* on these. Forced alignment is a speech processing technique that fits a given segmentation, in our case, the transcription, to a speech signal. Several aligners exist; for German, one of the most easily accessible and prominent ones, WebMAUS, is provided by the Bavarian Archive for Speech Signals (BAS), as part of their web services (Kisler, Reichel and Schiel 2017; Draxler, Harrington and Schiel 2017).[22]

We have been experimenting successfully with integrating WebMAUS into our workflow. However, we also found that it would be useful to have an alternative, as in many cases, data cannot be sent to web services such as those provided by the BAS, for three possible reasons. We report on our experiments with the BAS services, in particular. First, often the audio quality is insufficient for speech processing. Secondly, data may be too sensitive to transmit to external services which cannot guarantee encrypted storage; this applies to many of the corpora that can only be made accessible under restricted conditions. Thirdly, BAS web services often have problems if recordings are too long (where the limit may be as short as ten minutes) or contain certain features such as long pauses; both tend to occur in the data relevant to our current use case.

The processing performed in this step permits to estimate the alignment between temporal data (sound, video) and transcriptions relying on the graph(em)ic form of utterances, i.e. counting letters, or the canonical phonetic transcriptions provided by BAS web services, counting phone(me)s. Optionally, the canonical phonetic transcription can be added to the TEI-ISO document using the attribute `@phon` on `<w>`

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

[21] Unfortunately, we cannot give a general figure of accuracy for POS tagging in this context, but have to refer to general papers such as those by Schmid (1995) or Giesbrecht and Evert (2009). We would very much welcome the development of models for spoken language for languages other than German and French.

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

elements. As there is as yet no good TEI attribute for this[23], we use a non-standard attribute `@phon`, as for practical reasons, we try to avoid extra `<spanGrp>` elements for simple annotations.

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

Graph(eme)-based measures are useful because graph(eme)-to-phone(me) conversion is not available for every language, because of difficulties with language tagging ambiguity (see next paragraph) and because of the fact that timeline `<anchor>`s can occur even in words, and then it is impossible to determine the phone boundary. For instance, if an overlap (marked |) starts in the middle of the word “psycho|logist”, it is difficult to guess the correct breaking point. The algorithm regresses to counting letters instead.

A difficulty arises with respect to language handling, as for some languages, BAS web services use fully qualified locales as parameters. The service will do some adjustment to be able to transcribe (e.g., accept `ltz` and not just the full `ltz-LU` for Luxemburgish).

```
<timeline><when id="T0" interval="0.0s" since="T0"/>
  <when xml:id="B_2" interval="5.394s" since="T0"/>
  <when xml:id="E_2" interval="6.356" since="T0"/> ... </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> ...
```

Starting with `guess`, all steps are based on heuristics and NLP. Therefore, the results of these steps should be

4.7 Adressable elements (`identify` and `unidentify`)

There are two more services, which are only useful in specific cases. Occasionally, it is useful if all structural elements can be addessed with an `@xml:id` attribute. Hence, `identify` adds `@xml:id` attributes to all TEI elements that do not have one, and `@unidentify` removes such attributes whose form suggests they have been added by `identify`.

4.8 Integration with WebLicht: Parameters and an Optional Header

WebLicht (E. Hinrichs, M. Hinrichs and Zastrow 2010)[24] has proven successful, especially as a didactic and explorative environment for running webservices for linguistic annotation, and it is an important part of the CLARIN infrastructure. WebLicht's architecture is built around the pivot format TCF (see E. Hinrichs, M. Hinrichs and Zastrow 2010) which is currently in ins fifth version.[25]

The TEILicht services have been integrated into WebLicht and will be integrated into the Language Resource Switchboard (`https://switchboard.clarin.eu/`). The WebLicht team provided much help, and also a new input type for the plain text transcripts. The integration was not seamless, however. At this point in time, WebLicht's requirement to explicitly list all possible values for a given parameter poses problems for parameters with a large or continuous value set (such as languages, audio duration etc.).

Let us consider the case of language tags at some length. Of course, in the most common cases, a simple language code such as `de` or `nl-BE` may be sufficient. However, the language codes suggested by BCP 47, recommended by the TEI guidelines[26], are actually an open class and allow for impromptu tags like `de-DE-x-goethe` (example taken from BCP 47, page 10). Moreover, even offering a full list of languages that can be selected with two or three letter codes, and even more so offering to select several from this list is problematic. Therefore, the `guess` webservice was modified to accept four one-language identifiers `expected1` to `expected4`. These are used in addition to the list-valued `expected` parameter to restrict the search space of languages.

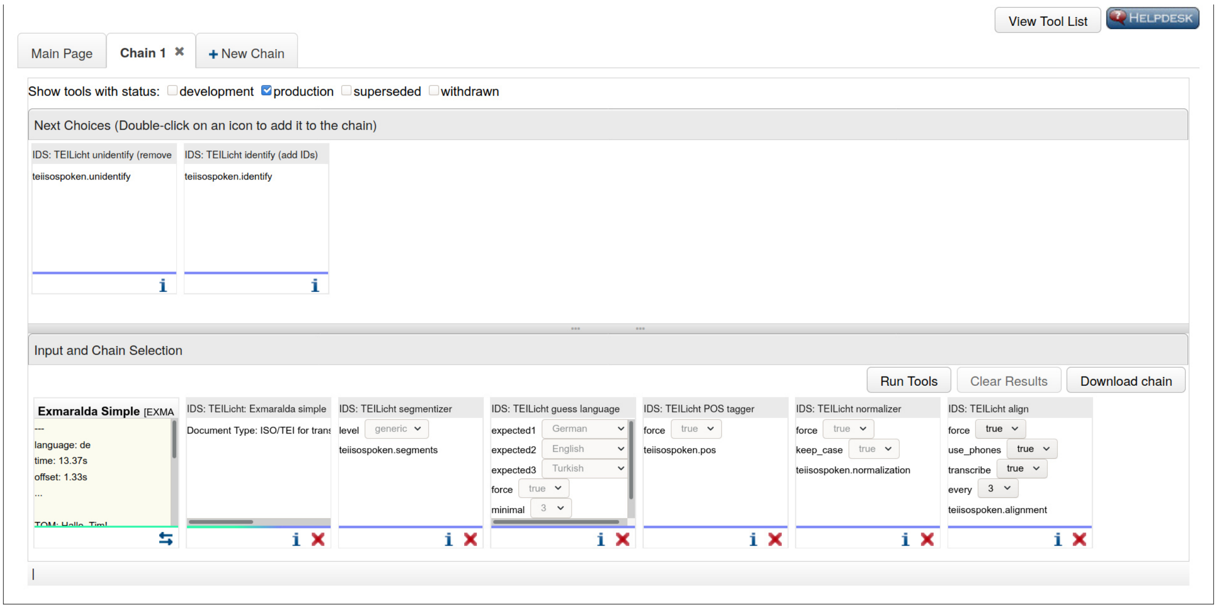
The only way to proceed with values such as positive integers (e.g., the `every` parameter of `align` for inserting a time anchor every n words) is to offer a snsible choice of values (e.g. 3, 5 and 10). For values such as positive floating point numbers such as the duration of a transcript or the `offset` parameter of the `align` service, it is not possible to find a good representation in WebLicht. However, these are important

[23] The integration of `@phon` has been submitted as a request to TEI.

[24] see `https://weblicht.sfs.uni-tuebingen.de/`

[25] The current specification is avaiblable at `https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format`.

[26] `https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teidata.language.html`



Show tools with status: □ development ☑ production □ superseded □ withdrawn

Next Choices (Double-click on an icon to add it to the chain)

IDS: TEILicht unidentify (remove
teiisospoken.unidentify

IDS: TEILicht identify (add IDs)
teiisospoken.identify

Input and Chain Selection

Exmaralda Simple [EXMA
---
language: de
time: 13.37s
offset: 1.33s
...
TOM: Hallo, Tim!

IDS: TEILicht: Exmaralda simple
Document Type: ISO/TEI for trans

IDS: TEILicht segmentizer
level generic
teiisospoken.segments

IDS: TEILicht guess language
expected1 German
expected2 English
expected3 Turkish
force true
minimal 3

IDS: TEILicht POS tagger
force true
teiisospoken.pos

IDS: TEILicht normalizer
force true
keep_case true
teiisospoken.normalization

IDS: TEILicht align
force true
use_phones true
transcribe true
every 3
teiisospoken.alignment

Figure 1: WebLicht integration, example chain combining most services in a sensible order

for the correct operation of the `align` service. To allow the processing, the plain text format was enhanced with an optional header where one can specify the main document language, the duration until end of the last utterance, and the offset of the first utterance can be specified, the example below. All specifications are optional, but an offset is only accepted if the duration is also specified.

```
---
lang: de
duration: 43s
offset: 0.0
---
```

4.9 Command line version and web services

The functionality of all web services is also available as a Java library and command line tool, see `https://github.com/Exmaralda-Org/teispeechtools/`. On bulk data, the command line tool is easier to use than the web services. Moreover, the command line invocation is generally free of privacy concerns, as no data are sent through the web.[27]

The commands for the command line tools have the same name as the services described above.

The parameters of the web service and the command line version generally have the same name, e.g. `--lang` (or `--language`) on the command line and `lang` in the web services, but with dashes swapped for underscores, e.g. `--minimal-length` (alternatively, `--minimal`) on the command line and `minimal_length` in the web service, and some shorter option names provided for the command line tool.

The following simulates an example run with the provided wrapper script `spindel.sh`. The `--indent` parameter causes the output XML file to be pretty-printed, mainly useful for debugging.

```
spindel.sh segmentize --lang=de -i 0-text2iso.xml --indent -o 1-segmentize.xml
spindel.sh guess --input=1-segmentize.xml --indent --output=2-guess.xml
spindel.sh normalize --input=2-guess.xml --indent --output=3-normalize.xml
spindel.sh pos --input=3-normalize.xml --indent --output=4-pos.xml
spindel.sh identify --input=4-pos.xml --indent --output=5-identify.xml
spindel.sh unidentify --input=5-identify.xml --indent --output=6-unidentify.xml
spindel.sh align -i 6-unidentify.xml --indent -o 7-align.xml --time 43 --every 5 -t
```

In the last step, `-t` causes transcription via the BAS web service to be added.

[27] But note that the `align` service may call the BAS transcription web service!

5 Conclusion

We have presented web services which implement a workflow for transcripts of spoken language which follow the TEI guidelines and in particular ISO 24624:2016 "Language resource management – Transcription of spoken language". These web services were illustrated with respect to a use case that occurs frequently in our daily work at the IDS. We hope to have shown that these web services are useful for a broader public, and form a useful addition to the CLARIN universe.

6 Outlook

The web services are currently available from IDS, and have been integrated into the CLARIN infrastructure, so that they can be found in the Virtual Language Observatory and can also be used in WebLicht.

For tagging, we shall have to evaluate whether it is useful to offer a direct choice of the tagger models for specific languages, as we now prefer models for spoken language where they are available, i.e. in the case of French and German, and use one of three models in the case of Portuguese.

It may also be worthwhile to test whether language detection with moving windows can be applied to longer utterances in a way that detects language shifts like code switching.

As regards alignment, we intend to evaluate pseudoalignment more than impressionistically, and we intend to evaluate further forced alignment tools.

In the long run, we will also be able to evaluate in which form such tools are best distributed. While WebLicht is very useful as a didactic showroom and can help to quickly and easily explore how a given tool works, it is probably not the tool of choice for batch operation in curation work on larger datasets. Standalone webservices may serve this purpose better, but can still bring a considerable overhead, or even constitute an obstacle when legal restrictions do not permit sending data over the internet. At least as long as data curation remains an expert job carried out in specialised data centres, plain command line tools as described in section sec. 4.9 are likely to remain a candidate for the most adequate option.

References

Barras, Claude et al. 2001. "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production". In: *Speech Communication* 33.1–2, pp. 5–22.

Betten, Anne, ed. 1995. *Sprachbewahrung nach der Emigration – Das Deutsch der 20er Jahre in Israel. Teil I: Transkripte und Tondokumente. unter Mitarbeit von Sigrid Graßl*. Phonai 42. Tübingen: Niemeyer.

Brinckmann, Caren et al. 2008. "German Today: a really extensive Corpus of Spoken Standard German". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2008/pdf/806_paper.pdf`.

Broeder, Daan et al. 2012. "CMDI: a component metadata infrastructure". In: *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*.

Clyne, Michael. 1981. *Deutsch als Muttersprache in Australien. Zur Ökologie einer Einwanderersprache. In Zusammenarbeit mit dem Centre for Migrant Studies, Monash University.* Wiesbaden: Franz Steiner.

CLARIN ERIC. 2019. *Component Metadata*. URL: `https://www.clarin.eu/content/component-metadata`.

Devran, Serap. 2017. *Deutsch-türkische Migration: Die Darstellung narrativer Identitäten von Studentinnen in Istanbul. Eine biografie- und interaktionsanalytische Pilotstudie*. amades. Mannheim: Institut für Deutsche Sprache.

Draxler, Christoph, Jonathan Harrington and Florian Schiel. 2017. "Towards the next generation of speech tools and corpora". In: *Computer Speech and Language* 46, pp. 175–178.

Giesbrecht, Eugenie and Stefan Evert. 2009. "Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus". In: *Web as Corpus Workshop (WAC5)*.

Hinrichs, Erhard, Marie Hinrichs and Thomas Zastrow. 2010. "WebLicht: Web-Based LRT Services for German". In: *Proceedings of the ACL 2010 System Demonstrations*. ACLDemos '10. Uppsala, Sweden: Association for Computational Linguistics, pp. 25–29.

ISO. 2016. *ISO 24624:2016 Language resource management – Transcription of spoken language*. Tech. rep. Genève: ISO.

Kipp, Sandra Joy. 2002. “German-English Bilingualism in the Western District of Victoria”. PhD thesis. Department of Linguistics and Applied Linguistics. The University of Melbourne.

Kisler, Thomas, Uwe D. Reichel and Florian Schiel. 2017. “Multilingual processing of speech via web services”. In: *Computer Speech and Language* 45, pp. 326–347.

Lehr, Ursula and Hans Thomae, eds. 1987. *Formen seelischen Alterns*. Stuttgart: Enke.

Lich, Glen E. and Michael Clyne. 1984. *Deutsch als Muttersprache in Australien: zur Ökologie einer Einwanderersprache. In Zusammenarbeit mit dem Centre for Migrant Studies, Monash University*. Wiesbaden: Franz Steiner.

Liégeois, Loïc et al. 2017. “Vers un format pivot commun pour la mutualisation, l’échange et l’analyse des corpus oraux”. In: *FLORAL*. Orléans, France.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum.

Schmid, Helmut. 1995. “Improvements In Part-of-Speech Tagging With an Application To German”. In: *In Proceedings of the ACL SIGDAT-Workshop*, pp. 47–50.

Schmidt, Thomas. 2011. “A TEI-based approach to standardising spoken language transcription”. In: *Journal of the Text Encoding Initiative* 1, pp. 1–22.

Schmidt, Thomas. 2012. “EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language”. In: *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC’12)*. Ed. by Thierry Declerck, Khalid Choukri and Nicoletta Calzolari. European Language Resources Association (ELRA), pp. 236–240.

Schmidt, Thomas. 2016. “Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project”. In: *Journal for Language Technology and Computational Linguistics* 31.1. Ed. by Marc Kupietz and Alexander Geyken, pp. 127–154.

Schmidt, Thomas. 2017. “DGD – die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim”. de. In: *Zeitschrift für germanistische Linguistik* 45.3. Ed. by Vilmos Ágel et al., pp. 451–463. URL: `http://nbn-resolving.de/urn:nbn:de:bsz:mh39-68145`.

Schmidt, Thomas. 2019. “Das Berliner Wendekorpus am Archiv für gesprochenes Deutsch”. In: *Sprechen im Umbruch. Zeitzeugen erzählen und argumentieren rund um den Fall der Mauer im Wendekorpus*. Ed. by Norbert Dittmar and Christine Paul. Mannheim: Leibniz-Institut für Deutsche Sprache (IDS), pp. 23–27.

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

Schmidt, Thomas and Wilfried Schütte. 2010. “FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction”. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Valletta, Malta: European Language Resources Association (ELRA). URL: `http://www.lrec-conf.org/proceedings/lrec2010/pdf/18_Paper.pdf`.

Schmidt, Thomas and Kai Wörner. 2014. “EXMARaLDA”. In: *The Oxford handbook of corpus phonology*. Ed. by Jacques Durand, Ulrike Gut and Gjert Kristoffersen. Oxford: Oxford University Press.

Sloetjes, Han. 2014. “ELAN: Multimedia Annotation Application”. In: *The Oxford handbook of corpus phonology*. Ed. by Jacques Durand, Ulrike Gut and Gjert Kristoffersen. Oxford: Oxford University Press.

TEI Consortium. 2019. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Tech. rep. Version 3.5.0. Last updated on 29th January 2019. TEI Consortium.

Thüne, Eva-Maria. 2019. *Gerettet. Berichte von Kindertransport und Auswanderung nach Großbritannien*. Berlin, Leipzig: Hentrich & Hentrich.

Verdonik, Darinka et al. 2013. “Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS”. In: *Language Resources and Evaluation* 47.4, pp. 1031–1048.

Westpfahl, Swantje. 2020. “POS-Tagging für Transkripte gesprochener Sprache. Entwicklung einer automatisierten Wortarten-Annotation am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)”. PhD thesis. Tübingen.

Wilkinson, Mark D. et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3, p. 160018. URL: `https://doi.org/10.1038/sdata.2016.18`.