Archimob Corpus
Release 1.0 documentation

CorpusLab*, URPP Language and Space

12 August 2016

*Heath Gordon (intern), Christof Bless (intern), Phillip Ströbel (assistant), Fatima Stadler (assistant), Noëmi Aepli (assistant), Tanja Samardžić (director)

Contents

1 Introduction

This documentation describes the first release of the Archimob corpus of Swiss German compiled at the University of Zurich. The corpus represents German varieties spoken on the territory of Switzerland. It is the first electronic resource containing long samples of transcribed text. The size of this version of the corpus is 528 381 tokens.[1]

This corpus is intended to be used for linguistic research (primarily to study morphosyntactic regional variation) and for natural language processing. This is why it is intended to be distributed in two ways. First, users can search the corpus through a corpus query engine. Second, an archive containing the transcriptions is available for download.

The corpus is free for research purposes. For commercial use, special contracts need to be made.

2 The Source and the Size of the Data

The original Archimob project was initiated by Frédéric Gonseth in 1998 and was conducted by the Archimob association.[2] It is a collaboration of historians and filmmakers in order to gather oral history data on the period of 1939–1945 in Switzerland. Their archive contains 555 recordings of interviews with Swiss citizens who witnessed the Second World War, covering topics such as political wrangling, daily life and even illicit love affairs during wartime. Out of these 555 recordings, 300 are in Swiss German. Each recording is produced with one informant using a semi-directive technique and usually is between 1h and 2h long. Informants come from all linguistic regions of Switzerland and represent both genders, different social backgrounds, and different political views.

The compilation of the present Archimob corpus started in 2004, when a collection of 52 VHS tapes was obtained from the Archimob association. The initial corpus composition was part of Janine Richner-Steiner's and Matthias Friedli's PhD projects, supervised by Elvira Glaser from the German Department of the University of Zurich. The Archimob data was used to investigate dialectal phenomena such as the varying position of the indefinite article in adverbially complemented noun phrases (Richner-Steiner, 2011) and comparative clauses in Swiss German (Friedli, 2012).[3] Richner-Steiner and Friedli selected interviews suitable for dialectal studies and established an inventory of the VHS tapes. In order to chose the material to be transcribed, the recordings were rated as category A, B, or C according to the linguistic representativeness of the speakers and the sound quality. Speakers who were not exposed to dialect/language contact are considered most representative for the place they live. 45 interviews with an A rating were then digitised at the German Department into the MP4 format.

[1] Two similar resources are currently available at the University of Zurich: The NOAH Corpus (Hollenstein and Aepli, 2014), consisting of roughly 115 000 tokens, taken from various written texts, and the Swiss SMS Corpus (Ueberwasser, 2009), consisting of 650 000 tokens, 41% of which are Swiss German. Furthermore a current project of the phonogram archives of the University of Zurich opts to digitise the collected and stored data, in order to build the PAZTeK corpus, in which queries on the dialectal material of different times and sources shall be possible (http://www.phonogrammarchiv.uzh.ch/en/projekte.html).

[2] Archimob (archives de la mobilisation): http://www.archimob.ch/.

[3] The selected archimob data is a part of and Friedli's and Richner-Steiner's Dialekt Korpus Zürich (DiaKoZ).

Phase	**1**	**2**	**3**
Duration	2006–2011	2011–2014	2015
Documents	16	7	11
Funding	Stiftung für wissenschaftliche Forschung UZH	Stiftung für wissenschaftliche Forschung UZH	Exchange with Spitch
Transcriber	Eva Peters	Patrick Mächler	Noëmi Aepli Alexandra Zoller
Software	Nisus Writer	FOLKER	EXMARAaLDA
Responsible	Elvira Glaser, Matthias Friedli	Elvira Glaser	Tanja Samardžić, Elvira Glaser

Table 1: The overview of the work on transcribing the Archimob recordings.

One video was excluded later because the rating was not appropriate. Further information on the history of the initial Archimob project and its integration in linguistic research can be found in (Friedli, 2012).

All the recordings of the category A, finally a total of 44 MP4 files, were selected to be included in the Archimob corpus. Not all interviews have been transcribed yet. The first release of the corpus contains 34 recordings transcribed with 15 540 tokens per recording on average. The remaining 10 recordings are planned to be be processed for the next release.

3 Transcription and Text Encoding

The interviews are transcribed in three phases, summarised in Table 1. In the first phase Eva Peters performed the transcription on 16 recordings without any specialised software. In the second phase, 7 documents were transcribed by Patrick Mächler, using the transcription tool FOLKER available from the IDS (Institute of the German Language) in Mannheim (Schmidt, 2012).[4] The work in the first two phases was done at the German Department, supervised by Elvira Glaser and Matthias Friedli and funded by the Stiftung für wissenschaftliche Forschung an der Universität Zürich. In 2014, the Archimob team is extended to include the newly founded URPP Language and Space CorpusLab. From that moment on, the work on the Archimob corpus is carried out jointly by the German Department and the CorpusLab. The third phase of transcriptions was enabled by an exchange agreement between the extended team and the private start-up Spitch, based in Zurich. Spitch provided funding for transcribing 11 new documents in exchange for the existing transcriptions. The documents covered by this agreement represent the Zurich variety and a number of similar dialects. The transcriptions in the third phase were done by Noëmi Aepli and Alexandra Zoller using the tool EXMARaLDA, also available from the IDS in Mannheim (Schmidt, 2012).[5] The work was coordinated by Tanja Samardžić and supervised by Elvira Glaser and Patrick Mächler.

[4] http://agd.ids-mannheim.de/folker_en.shtml
[5] http://www.exmaralda.org/

3.1 Writing

There is no widely spread convention for writing Swiss German. We use the writing system "Schwyzertütschi Dialäktschrift" proposed by Dieth (Dieth, 1986) developed to provide some guidance on how to write in a Swiss German dialect. The transcription is expected to show the phonetic properties of the variety but in a way that is legible for everybody who is familiar with standard German text reading (Dieth, 1986, 10). Dieth's system, which is originally phonemic, can be implemented in different ways depending on how differentiated the phonetic qualities are to be expressed. Although it is the objective to keep track of the pronunciation, Dieth's transcription method is orthographic and partially adapted to spelling habits in standard German. Therefore it does not provide the same precision and explicitness as phonetic transcription methods do.

We do not use the full power of phonemic distinctions available in the Dieth script. The practice in using Dieth's system changed over the time, so that more distinctions concerning the openness of vowels were made in the first phase than in the later phases. The precise decision on how to apply the Dieth script are documented in the project specific transcription guidelines. For the latest version see (Mächler, 2015b).

The grapheme inventory in the Dieth's script is always related to the dialect and its phonetic properties, so that, for example, the grapheme <e> stands for different vowel qualities, [e], [ɛ] or [ə], depending on the dialect, the accentuation of the syllable and – to substantial degree – also to the dialectal background of the transcriber. The guidelines needed to be adapted over the time because the variety-dependent transcription system was hard to implement in a consistent way. The most important changes concern vowel quality (1), capitalisation and the degree of adaption to standard German orthography (2). The examples are taken from interviews with Zurich German speakers, where the transcription differences are not explained by the dialect variation but on slightly different spelling methods.[6]

It is worth noting that the transcription is focused on the audio source, so that gestures and other visual elements are described minimally.

(1) Phase 1: *èèr* vs. phase 3: *er* (std. *er*, engl. 'he')

(2) Phase 1: *wän* vs. phase 3: *wänn* (std. *wenn*, engl. 'if')

3.2 Speech-to-text alignment

The transcriptions are aligned with the sound recordings at the level of a transcription unit, usually of the length between 4 and 8 seconds. These units are manually formed by transcribers. Such alignment is part of the output of specialised tools like FOLKER and EXMARaLDA. Since no specialised tool was used in the first phase (see Table 1), the transcriptions from this phase needed to be segmented and aligned subsequently. The segmentation of these documents was done in two steps. In the first step, we produce automatically approximative segmentations and format the texts as an EXMARaLDA (.exb) file. In the second step, we import the approximative segmentations into EXMARaLDA and correct the unit borders manually. This alignment was performed as part

[6]Speaker ID's: 1143, 1188.

of the collaboration with Spitch in the case of the Zurich German recordings and by the CorpusLab for the remaining recordings.

Some of the first phase transcriptions were previously automatically aligned at the level of word using the tool WebMAUS available from the University of Munich (Kisler et al., 2012). This work was done by Hanna Ruch (URPP Language and Space), Anne Göhring and Alexandra Bünzli (Institute of Computational Linguistics). The funding was granted by a 2014 ZüKL KLIP project. For those documents for which it existed, we used this alignment as a pre-processing: the automatically aligned words are automatically grouped into our target segments before importing the text into EXMARaLDA for manual correction. When the pre-processing was not available, we first segmented the transcribed text automatically based on some indicators of pauses in the transcriptions and then corrected the segmentation and alignment manually using EXMARaLDA.

To unify the transcription formats for further processing, we also converted the FOLKER output files into the EXMARaLDA format using the EXAKT tool.[7]

3.3 Text format

The final corpus format is an instance of XML based on the Text Encoding Initiative (TEI) recommendations. We follow the TEI recommendations whenever it is possible and add specific elements only for the cases not explicitly covered by TEI. The final XML format is verticalised text where each word is an XML element so that word-level annotation can be encoded in the attributes. Since each transcription phase resulted in a different format, we perform several different conversions to arrive at the final format designed by Noëmi Aepli, Phillip Ströbel and Tanja Samardžić (CorpusLab). The first phase transcriptions needed to be converted from the output of Nisus Writer into an XML format, which was done by Anne Göhring. The other conversions, performed by Noëmi Aepli and Phillip Ströbel, concerned different XML formats.

In addition to unifying the format, the final corpus format includes a metadata scheme, designed by Heath Gordon (CorpusLab intern), Phillip Ströbel, Noëmi Aepli and Tanja Samardžić. Furthermore a unified sound alignment encoding, based on EXMARaLDA, was implemented by Phillip Ströbel. The structure of the current XML format is described in more detail below.

4 Annotation

The corpus contains word level annotation where we specify for each token whether it is a word or some other conversational element (pauses or interruptions). Words are assigned a normalised form and a part-of-speech tag. More specialised annotations like the categorisation of proper names or particular dialectal features for instance are planned to be provided in extra XML files pointing to the corresponding token in the basic document in next releases.

[7] `http://www.exmaralda.org/tool/exakt/`

4.1 Normalisation

Due to the properties of variation and the absence of standardisation on the level of dialectal data, the usual lemmatisation practice in natural language processing is not satisfactory in the case of non-standard varieties. During an exploratory workshop on normalisation, organised by Agnes Kolmer and funded by the Swiss National Science Foundation (SNSF) in 2011, a group of linguists of the University of Zurich discussed normalisation practice in historical linguistics and non-standard varieties. The normalisation implemented in this corpus is based on these discussions.

Variation in written Swiss German is generally observed at two levels. First, a lexical unit that can be identified as "the same word" is pronounced, and therefore also written, in a different way in different regions of Switzerland (3). Second, a lexical unit that can be considered phonetically invariant (within a region) is written in a different way on different occasions (4). The two types of variation combined result in a great number of potential variants that need to be reduced to a single form in order to establish identity between words that are felt to be the same across variants.

(3) [χriɛg] vs. [kriag] ('war')

(4) <gsii> vs. <xii> ('been')[8]

Normalisation of Swiss German usually resembles standard German. There are, however, many possible approaches to this task, depending on how close the normalised form is to the standard form. With the goal to represent the local varieties as accurately as possible, we opt for a normalised representation of a Swiss German construction, respecting etymological relations and avoiding the known sources of inconsistency. We adopt standard German spelling where possible, but we do not translate words specific to Swiss variants. Our general approach is to distinguish between two cases:

1. The word is transformed into a standard German version following the etymological principle

2. In cases where there is no standard German equivalent, a normalised form of the dialectal word is implemented

The general procedure of the normalisation is to transform every morpheme of a Swiss German word into a normalised standard German version, following an etymological principle. This means that every morpheme has to be normalised with the etymologically most likely correspondent morpheme in the standard variety, if it exists. However, discrepancies in word formation (5), inflection or in the meaning of the word (6) are not represented. Morphosyntactic features in Swiss German lexemes that are not realised in standard German, on the other hand, are transformed into morphologically transparent normalisations. This procedure leads sometimes to more complex normalised forms compared to the corresponding standard German lexemes (7).

[8]Examples are taken from the sms4science corpus: The access is granted after registration on https://sms.linguistik.uzh.ch/bin/view/Main/WebHome.

(5) *d Manne* 'the men' → *die Männer*
tagreis 'daytrip' → *Tagesreise*
Not normalised to *Mannen* or *Tagreise*.

(6) *züglet* 'moved' (German *umgezogen*) → *gezügelt*
In standard German *gezügelt* means 'bitted'.

(7) *dur**e*** 'through' (German *durch* + direction) → *durchhin*

In addition to these general rules a number of case-based decisions needed to be taken. The information about the choices made in particular cases is available in the guidelines written by Patrick Mächler (Mächler, 2015a).

The second procedure to normalise the data is applied when standard German misses an etymological equivalent. In this case the normalisation needs to be constructed on a non standard base (8). The normalised form should represent several dialectal variants of a genuine Swiss German lexeme. The construction of a representing form is a hypothetical standard German form on the base of a regular correspondence to Middle High German, mostly according to the lemma of the *Idiotikon* (Staub et al., 1881).[9] These pro forms are registered in a list as part of the normalisation guidelines.

(8) *gumpe* "jump" (German "springen") → *gumpen*
niene "nowhere" (German "nirgendwo", "nirgends") → *niener*
tätschts "pop" (German "knallen") → *tätschen*

It is important to note that the normalised data is not intended to be known by human users. It is a hidden annotation layer used only for automatic processing. The users are expected to formulate queries and the results are presented in a form of original writing (keeping the original inconsistency). This allows us to choose arbitrary representations, which users would find artificial and hard to adopt.

4.1.1 Manual normalisation: Phase 1 & 2

The normalisation was performed in three phases, described in more detail in the following sections. The two predominantly manual procedures, are summarised in Table 2.

In the first phase, two transcribed documents were normalised manually by Patrick Mächler and Franziska Schmid using the tool VARD2 (Baron and Rayson, 2008). The tool was initially tested by Alexandra Bünzli. Anne Göhring was responsible for the installation and the further management of the application. The normalisation process was supervised by Elvira Glaser and Agnes Kolmer. During this phase the approach to the normalisation was elaborated and documented in the guidelines.

In the second phase, four documents were normalised by Noemi Graf and Mike Lingg, using the tool SGT (Ruef and Ueberwasser, 2013), which was adapted to the project's needs by Anne Göhring and later on installed and managed by Alexandra Bünzli. The work was funded by a ZüKL KLIP project, coordinated by Tanja Samardžić, and supervised by Patrick Mächler and Elvira Glaser.

[9]The entries in the *Idiotikon* are normalised, early attested Swiss German forms or (reconstructions) of Middle High German ancestors (Staub et al., 1881, XIV).

Phase	**1**	**2**
Duration	2011-2013	2014
Documents	2	4
Funding	Multilingual Text Analysis at UZH	ZüKL KLIP
Collaborator	Patrick Mächler, Franziska Schmid	Noemi Graf, Mike Lingg
Software	VARD2	SGT
Responsible	Elvira Glaser, Agnes Kolmer	Tanja Samardžić, Elvira Glaser, Patrick Mächler

Table 2: The manual procedure of the Archimob transcriptions.

The documents normalised in the first and the second phase followed the same general approach. There are, however, some differences due to some changes in the guidelines and the different software. For example, in the second phase the collaborators used upper and lower case characters according to standard German rules, while the normalisation was all in lower case in the beginning. Since standard German capitalisation rules are judged unnecessary for this corpus, it was decided later on to put the normalisation all into lower case. Another difference between the two manual normalisation stages is that the first phase contains special marks attached to the normalised form to signalise certain properties of the token, for example the prepositional dative marker that does not exist in standard German. Such information is not agglutinated in the second phase. SGT provides check boxes to mark various special cases, so that this information is separated from the normalisations themselves. This method allows to annotate named entities, word interruptions, unclear content etc. properly as XML elements or attributes (see 5.3).

4.1.2 Automatic normalisation: Phase 3

In the third phase six manually normalised documents are used to train a system for automatic normalisation. During a first experimental round, in 2015, the system, developed by Tanja Samardžić and Yves Scherrer reaches an accuracy score of 77.28%. More information about the methods can be found in the paper *Normalising orthographic and dialectal variants for the automatic processing of Swiss German* (Samardžić et al., 2015).

While evaluating the system performance in the first experimental round, it turned out that the normalisations by different annotators show considerable inconsistency. These discrepancies are the result of adaptions of the normalisation rules, not explicitly treated cases in the guidelines or simply individual deviations from them. The detected inconsistencies are then corrected both, automatically and manually by Fatima Stadler and Yves Scherrer. With the consolidation of the training set, the performance of the automatic normalisation enhances to an accuracy of 84.13% as measured in a cross-validation on the 6 manually normalised documents (Samardžić et al., 2016). In addition to the cross-validation, this version of the system was evaluated on three new documents arbitrarily selected from the remaining documents. These documents were processed by Yves Scherrer using the system trained on the whole manually normalised set and then manually corrected by Philip Ströbel and Fatima Stadler. The accuracy was 87.58% in the document in the variety closest to those represented in the training set and around 78% in the two other doc-

uments, more distant from the training set. To asses whether an increased training set improves the performance on distant varieties, two of the manually corrected documents, one close and one distant, were added to the training set. The enriched model is then used to evaluate the system on the other distant documents. Including additional data did not bring improvements, as the accuracy remained at around 78%. The details of this evaluation are reported in (Samardžić et al., 2016).

The performance of the automatic system is further improved by Yves Scherrer in collaboration with Nikola Ljubešić to the accuracy score of 90.46%, as reported in the paper *Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation* (Scherrer and Ljubešić, 2016).

All the documents in this release that are not manually normalised are processed using the version of the system reported in (Scherrer and Ljubešić, 2016). An improvement of the normalisation system is planned for the next edition of the corpus.

Normalisation process	**Files**
Manually normalised	1007, 1048, 1063, 1143, 1198, 1270
Manually corrected	1142, 1212, 1261

Table 3: Overview over the normalisation phases. 6 files were manually normalised, 3 files were automatically normalised and manually corrected. The rest was processed only automatically, without manual correction.

4.2 Part-of-Speech Tagging

Part-of-speech tags enable formal searches abstracting away from lexical units. This annotation is thus a crucial step in making a corpus accessible for linguistic research through improved search facilities.

In approaching this task, we rely on previously developed tools and resources that need to be adapted for our corpus. To be able to asses the performance of the tools, we first created a test set with manually annotated part-of-speech tags. This set contains a subset of the Archimob corpus, consisting of 10 169 tokens in 1742 utterances. The annotation was performed 2015 by Noëmi Aepli and coordinated by Tanja Samardžić.

The annotation guidelines follow those by (Hollenstein and Aepli, 2014). The tagset is the Stuttgart-Tübingen-Tagset (STTS) (Thielen et al., 1999) with some extensions to account for specific phenomena observed in Swiss German. An example of a tag which is exclusive to Swiss German is PTKINF used to tag the *go* in a sentence such as *Ich gang go poschte* 'I am going shopping' (German 'Ich gehe einkaufen'). Another reason why special tags are needed is that words are often written together in Swiss German when they would be written separately in standard German. For example, the English phrase *it is* under certain conditions corresponds to standard German *ist es*, which becomes *isches* in Swiss German. To tag *isches* as a single word, the solution is to append a '+' to the tag for the first word in the concatenation. All PoS tags with a '+' at the end mark Swiss German's concatenated forms. It is important to note that the '+' sign can appear in a concatenation too, where more than two lexemes are be merged:

(9) *hetmese* (German: 'hat man sie', eng. 'someone has ... her/them'): VAFIN+PIS+PPER

Several part-of-speech tagging experiments were performed by Tanja Samardžić and Yves Scherrer to find the best starting point for annotating the corpus. All tagging experiments are carried out with BTagger (Gesmundo and Samardžić, 2012), which has shown good performance on smaller training sets. Two models trained on similar data set are tried as the starting models:

- TüBa-D/S (Dipper et al., 2013): transcribed dialogues conducted in standard German (360 000 tokens in 38 000 utterances)
- NOAH's Corpus of Swiss German Dialects (Hollenstein and Aepli, 2014): Swiss German texts from various written sources (73 000 tokens) (not normalised)

TüBa-D/S model achieved an accuracy of only 70.68% when it was tested on the normalised version of the transcriptions. The NOAH model performed with the accuracy of 73.09% when tested on the original (not normalised transcirptions). Both results were achieved when the punctuation in the training corpora removed. The tagger was adapted by Noëmi Aepli, who manually corrected the automatic tagging in 4 ArchiMob documents (1048, 1063, 1198 and 1270) and then retrained the system adding the ArchiMob gold data to the NOAH training set. After the adaptation, the accuracy reached 90.09%. More details and the discussion can be found in *Archimob — A corpus of Spoken Swiss German* (Samardžić et al., 2016).

All the documents in this release, except the four manually corrected documents, are PoS tagged using the adapted tagger. An improvement of the normalisation system is planned for the next edition of the corpus.

5 XML Specifications

The corpus encoding stays as close as possible to the current TEI standards.[10] The XML schema of the Archimob corpus differs from the TEI schema only in the cases where TEI does not provide the needed tag or attribute. This is the case, for instance, with the attribute that we used to mark the normalisation. The data are stored in three types of files:

- **content files**, that contains the text of transcriptions marked with XML.
- a **media file**, that contains the alignment between transcribed text and the corresponding video documents.[11]
- a **speaker file** including the socio-demographic information about the informants (region/dialect, age, gender, occupation) and the information about the speakers' roles in the conversation (interviewer, interviewee).[12]

For most uses, the content files are sufficient. If any timestamps or information about the speakers are needed, this information can be included from

[10] http://www.tei-c.org/index.xml.
[11] Video documents are shared on request.
[12] Note that the roles of the interviewer and the interviewee are constant in our corpus.

```
<u start="media_pointers#d1007-T191" xml:id="d1007-u95" who="person_db#EJos1007">
      <w normalised="das ist" tag="PDS+" xml:id="d1007-u95-w1">dasch</w>
      <w normalised="glaube" tag="VVFIN" xml:id="d1007-u95-w2">glòüb</w>
      <w normalised="ich" tag="PPER" xml:id="d1007-u95-w3">ich</w>
      <vocal>
        <desc xml:id="d1007-u95-w4">eh</desc>
      </vocal>
      <w normalised="mein" tag="PPOSAT" xml:id="d1007-u95-w5">mi</w>
      <w normalised="bruder" tag="NN" xml:id="d1007-u95-w6">brieder</w>
      <w normalised="der" tag="ART" xml:id="d1007-u95-w7">de</w>
      <w normalised="sagt" tag="VVFIN" xml:id="d1007-u95-w8">sait</w>
      <w normalised="er" tag="PPER" xml:id="d1007-u95-w9">er</w>
      <w normalised="ja" tag="ADV" xml:id="d1007-u95-w10">ja</w>
</u>
```

Figure 1: An illustration of the content file format.

the corresponding files. We currently do not use any mechanism for automatic inclusion of the sound alignment and speaker meta-data into the content files.

In order to guarantee the validity, schema-compliance and consistency of the produced XML corpus files, each file is validated against a schema stored in the file `schema\_release\_1.xsd`.

5.1 File names

The transcriptions were previously stored in documents denominated after the name of the interviewee. This made identifying the recordings hard, as sometimes the first name was used, sometimes the last name, sometimes both and in different order. We renamed all the documents (both the transcriptions and the corresponding videos) according to the interview IDs of the ArchiMob project database.

The documents transcribed in the second phase were split into two or three parts. To keep the correspondence between the transcriptions and the videos, we kept these files as they were split, adding a suffix to the name of the file indicating which part of the recording it contains. For instance, the document 1082 was split into three parts, which results in three .xml files "1082_1.xml", "1082_2.xml", "1082_3.xml". The corresponding video and sound files have the same name with the extension depending of the format (.wav or .mp4).

One recording transcribed in the first phase, the document 1075, was identified as 1057 in the first XML conversion (see above). We set the identifier back to 1075, which is consistent with the codes in the Archimob database.

5.2 Segmentation and Tokenisation

Segmentation in a transcription of spoken language is very different from written text segmentation since paragraph and sentence boundaries are not marked. We opt for the segmentation that results from the manual transcription with the software EXMARaLDA. Each segmentation unit corresponds to the element "event" in the software output format (.exb files), illustrated here with two segmentation units:

```
<event start="T1650" end="T1652">wie / wie was händ sii </event>
```

```
<event start="T1652" end="T1654">deet erläbt </event>
```

The EXMARaLDA events correspond to the text-to-speech alignment units described in Section 3.2, that is to the points where the transcriber stops the video to write down the transcription. The attributes `start` and `end`, pointing to the time stamps, are used for encoding text-to-speech alignment.

The transcriptions, segmented and aligned to the sound source as described in Section 3.2, are converted from the .exb format into the final XML format.

The content files of the corpus are segmented into utterances, tagged as `<u>`, which correspond to the transcription units. They include references to the speaker and the media file specified as attributes. Note that we do not try to mark sentences boundaries. If this level of segmentation turns out to be useful, it is possible to add it in a later stage.

The final format of the corpus is built with a Python script, written by Noëmi Aepli with the filename transc_to_xml_na.py.[13] To run the script, the source files must lie in their respective folders regarding the transcription phase. To process the files and to create the XML files with the structure described in this documentation, the following command has to be executed from a terminal.

```
python transc_to_xml_na.py -d /PATH/TO/DATA/FOLDER -o /PATH/TO/OUTPUT/FOLDER
```

While segmentation is relatively hard to perform in spoken compared to written language, tokenisation is rather straightforward. As no punctuation is used in transcriptions, we take any string of characters between white spaces to be a token. Figure 1 shows an example of an utterance in our corpus. The XML elements, attributes and values are commented in more detail in Section 5.3.

5.3 Attributes and values in the content files

5.3.1 TEI Header

Detailed information about the corpus release, the corresponding video sources, etc. are given in the header of each document as in the example below.

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Transcription 1007</title>
    </titleStmt>
    <publicationStmt>
      <publisher>University of Zurich</publisher>
      <distributor>CorpusLab @ UFSP Sprache und Raum</distributor>
      <pubPlace>Zurich, Switzerland</pubPlace>
      <date>June 2016, Release 1.0</date>
    </publicationStmt>
    <sourceDesc>
      <recordingStmt>
        <recording type="video" xml:id="d1007">
          <respStmt>
            <name>ArchiMob Association</name>
```

[13] Not shared with the corpus, but available on request.

```
            <resp>www.archimob.ch/d/archimob.html</resp>
          </respStmt>
        </recording>
      </recordingStmt>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

The `<fileDesc>` element contains all the bibliographic information about the file which is transcribed. The `<titleStmt>` tag includes the title of the corpus file, in our case the identification of the transcription. The element `<publicationStmt>` specifies the details about the corpus publication, such as who published it, who distributes it, where and when it was published. The information about the data source is given in the element `<sourceDesc>`, where we specify that the data are originally collected as video recordings by the ArchiMob association. The element `<respStmt>` specifies that the Archimob association is the holder of the intellectual content.

5.3.2 Utterance

Each utterance element `<u>` has three attributes: `start`, `xml:id` and `who`.

start The value of the `start` attribute is a unique time ID that points to an absolute instant of time in the media file named media_pointers.xml.

```
<u start="media_pointers#d1007-T191" xml:id="d1007-u95" who="person_db#EJos1007">
```

In this example, the start of utterance 95 can be found in the media file with the ID d1007-T191.

xml:id This is a unique ID belonging to each utterance. The part in front of the hyphen is the document name, the part after it is the counter of the utterance. This is the 95^{th} utterance of the document 1007:

```
xml:id="d1007-u95
```

who The `who` attribute refers to the speaker of the utterance. Each interviewee has a unique ID in the content file, which points to his or her entrance in the separately stored document person_file.xml, where the informants' metadatas can be found. Concerning the interviewers, at this point, there is no further data available. The value of the attribute is always set to `"interviewer"` when he or she is taking the parole. When there are other speakers identified in the interview, their utterances are signalised by the label `"otherPerson"`.

5.3.3 Word

Each utterance is predominantly made up of `<w>` elements (words). The text of the element is the actual transcribed text as given to us by the transcribers. Each `<w>` element has 3 attributes: `normalised`, `tag` and `xml:id`.

xml:id Each word can be unambiguously identified by a unique ID. In the following example the ID specifies that it is the first word (or token) of the 95^{th} utterance of transcription document 1007.

```
xml:id="d1007-u95-w1"
```

normalised This attribute's value is the normalised form of the transcribed word as described in section 4.1.

tag This element indicates the part of speech tag of the word as described in section 4.2.

5.3.4 Gaps and Incertitude in Transcription

Some contents of the speech have not been clear enough to transcribe. In the cases where the transcribers were able to guess, they transcribed a word the way they heard it. On other occasions, the speech was unintelligible, so the collaborators were not capable of transcribing anything. Both cases involve material that is probably a linguistic content. Therefore these units, nested as `<gap>` or `<unclear>` tags, have an `xml:id` with a `w` suffix that is further counted to track on their position, even if these elements are not treated as `<w>` elements themselves.

<**gap**> Unintelligible words are not transcribable. The value of the element's attribute "type" is always `"untintelligible"` and the text is signalised by ellipsis `"..."`. An example of a gap element is given here:

```
<gap reason="unintelligible" xml:id="d1007-u108-w1">...</gap>
```

<**unclear**> This element is used, when a transcriber is not entirely sure about what he or she has heard. A suggested transcription for such ambiguous cases appears in the XML wrapped in an `<unclear>` tag. An `<unclear>` element can contain anything that an `<u>` tag can contain. For example:

```
<unclear>
    <w normalised="wir" tag="PPER" xml:id="d1007-u207-w10">mer</w>
    <w normalised="haben ihn" tag="VAFIN+" xml:id="d1007-u207-w11">hendne</w>
</unclear>
```

5.3.5 Non-linguistic Units

Some things were notated in the course of transcription that are not necessarily words. To mark such non-linguistic units we use the elements `<pause>`, `<del>`, `<vocal>`, `<kinesic>`, `<incident>` and `<other>`.

<**pause**> Pauses are marked by an empty element. For example, if a speaker says *gsii* then pauses shortly and continues with *guet*, this would be represented in the following way:

```
<w normalised="gewesen" tag="VAPP" xml:id="d1007-u234-w4">gsii</w>
<pause xml:id="d1007-u234-w5"/>
<w normalised="gut" tag="ADJD" xml:id="d1007-u234-w6">guet</w>
```

Note that a pause does not count as a word, neither do hesitations, comments, etc.

<**vocal**> If the utterance contains vocalised, but non-linguistic content, as hesitations, coughing, laughing etc., then it is marked like this:

```
<vocal>
    <desc xml:id="d1007-u238-w1">eh</desc>
</vocal>
```

The description tag `<desc>...</desc>` is a wrapper occurring in different kinds of elements (vocal, kinesic, incident, other) to specify their content. Here it contains a standardised hesitation marker.

<**del**> Sometimes, speakers stumble over their own words. An interrupted word is annotated as deletion in a `<del>` element:

```
<del type="truncation" xml:id="d1007-u269-w7">hundertvierz/</del>
```

The text of the element contains the truncated word with a slash to mark the interruption. The element provides also an attribute, `type="truncation"`, to indicate explicitly that the content of this element is fragmentary.[14]

<**kinesic**> On some occasions, the transcribers add comments concerning some specific behaviour of the speakers (e. g. body movement). Those comments are wrapped by the element `<kinesic>`, and specified in the `<desc>` tag. In the example below the interviewee gestures a lot, causing some noise in the recording because the attached microphone is moved by this behaviour (exact translation: 'moves, sizzling noise'). The transcribers noted this non-verbal information in square brackets.

```
<kinesic>
  <desc>[bewegt sich, knistert]</desc>
</kinesic>
```

<**incident**> Further information concerning the recording situation, for instance the changing of the recording tape or sounds from other sources are stored in the `<incident>` tag, and explained in the description element. The example shows a situation where a barking dog is intervening in the conversation (exact translation: 'barking'):

```
<incident>
  <desc>[hundegebell]</desc>
</incident>
```

[14] In our corpus this is redundant information but in the TEI standard, the <del> tag is used more broadly and the type attribute then can be useful. We prefer to be explicit and reinforce this information in our corpus than suppress it.

<**other**> The transcriptions contain some additional comments which cannot be classified as any of the listed non-linguistic categories. For example, the transcribers sometimes specify to whom the speaker is talking, or they comment on the quality of the sound in the recording, or try to guess what is said in the recording. Such comments are tagged with `<other>`. In the following example, the transcriber adds a description of the interviewees' simulation of motor noises (exact translation: 'imitates engine noise'):

```
<other>
  <desc>[ahmt motorengeräusch nach]</desc>
</other>
```

5.4 media_pointers.xml

All `<u>` (utterance) elements in the content files have a unique media attribute which points to a specific point in time listed in the media file `media\_pointers.xml`. We illustrate the structure of the media file with the following example:

```
<MediaPointers>
    <media docid="d1007">
      <timepoint absolute-time="5.253330269638217" xml:id="d1007-T0"/>
      <timepoint absolute-time="6.793329371524548" xml:id="d1007-T1"/>
      <timepoint absolute-time="9.186661309088153" xml:id="d1007-T2"/>
      <timepoint absolute-time="10.79332703876177" xml:id="d1007-T3"/>
```

The body of the file is wrapped in a `<MediaPointers>` element which consists of the elements `<media>`. Each `<media>` element corresponds to one document identified with the value of the `docid` attribute. This element consists of the elements `<timepoint>` with two attributes: the value of `absolute-time` refers to the time in the recording where the content file element with the corresponding `xml:id` starts.

5.5 person_file.xml

The information regarding the interviewees is stored in a separate file pointed to from the content files. Here is an example of the structure:

```
<body>
 <listPerson>
  <person xml:id="EJos1007" sex="f">
   <persName>Josy E***</persName>
   <birth when="-1912-01-26">26.01.1912</birth>
   <occupation>Haushaltsgehilfin</occupation>
   <residence>Stans, NW</residence>
  </person>
 </listPerson>
</body>
```

The element `<body>` contains `<listPerson>`, which contains the elements `<person>` where the information is stored. The value of the attribute `xml:id` matches the content file element produced by the given speaker (the corresponding attribute in the content files is `who`). We do not disclose the identity of the interviewees,

following the practice established by the ArchiMob association. The information stored in this files comes primarily from a sheet produced during the ranking of the videos as described above. The socio-demographic information is also available in the database of the ArchiMob project accessible online.[15]

[15] http://www.archimob.ch/arc/db/.

References

Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University.

Dieth, E. (1986). *Schwyzertütschi Dialäktschrift*. Sauerländer, Aarau, 2 edition.

Dipper, S., Lüdeling, A., and Reznicek, M. (2013). NoSta-D: A corpus of German non-standard varieties. In Zampieri, M. and Diwersy, S., editors, *Non-Standard Data Sources in Corpus-Based Research*, number 5 in ZSM-Studien, pages 69–76. Shaker.

Friedli, M. (2012). *Der Komparativanschluss im Schweizerdeutschen: Arealität, Variation und Wandel*. PhD thesis, Universität Zürich.

Gesmundo, A. and Samardžić, T. (2012). Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 368–372, Jeju Island, Korea. Association for Computational Linguistics.

Hollenstein, N. and Aepli, N. (2014). Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, COLING 2014, Dublin, Ireland. Association for Computational Linguistics.

Kisler, T., Schiel, F., and Sloetjes, H. (2012). Signal processing via web services: the use case webmaus. In *Proceedings Digital Humanities 2012, Hamburg, Germany*, pages 30–34, Hamburg.

Mächler, P. (2015a). Kommentar zur Normalisierung der ArchiMob-Texte 1007 und 1048, German Department, University of Zurich. Msc, German Department, University of Zurich.

Mächler, P. (2015b). Transkriptionsleitfaden fürs Schweizerdeutsche, msc., German Department, University of Zurich. Msc, German Department, University of Zurich.

Richner-Steiner, J. (2011). *"E ganz e liebi Frau". Zu den Stellungsvarianten des indefiniten Artikels in der adverbiell erweiterten Nominalphrase im Schweizerdeutschen. Eine dialektologische Untersuchung mit quantitativ-geographischem Fokus*. PhD thesis, Universität Zürich.

Ruef, B. and Ueberwasser, S. (2013). The taming of a dialect: Interlinear glossing of Swiss German text messages. In Zampieri, M. and Diwersy, S., editors, *Non-standard Data Sources in Corpus-based Research*, pages 61–68, Aachen.

Samardžić, T., Scherrer, Y., and Glaser, E. (2015). Normalising orthographic and dialectal variants for the automatic processing of swiss german. In *Proceedings of The 4th Biennial Workshop on Less-Resourced Languages*. European Language Resources Association (ELRA).

Samardžić, T., Scherrer, Y., and Glaser, E. (2016). ArchiMob – A corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož (Slovenia). European Language Resources Association (ELRA).

Scherrer, Y. and Ljubešić, N. (2016). Automatic normalisation of the swiss german archimob corpus using character-level machine translation. In *Proceedings of Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, Bochum, Germany.

Schmidt, T. (2012). EXMARaLDA and the FOLK tools. In *Proceedings of Language Resources and Evaluation (LREC 2012)*, Istanbul. European Language Resources Association (ELRA).

Staub, F., Tobler, L., Bachmann, A., Gröger, O., Wanner, H., and Dalcher, P., editors (1881). *Schweizerisches Idiotikon: Wörterbuch der schweizerdeutschen Sprache*. Huber, Frauenfeld.

Thielen, C., Schiller, A., Teufel, S., and Stöckert, C. (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, University of Stuttgart and University of Tübingen.

Ueberwasser, S. (2009). The swiss sms corpus, documentation, facts and figures.

A The list of collaborators

Noëmi Aepli (UZH, CorpusLab and Spitch)

Henning Beywl (UZH, CorpusLab)

Christof Bless (UZH, CorpusLab)

Alexandra Bünzli (UZH, ZüKL)

Fran Campillo (Spitch)

Matthias Friedli (UZH, German Dept)

Elvira Glaser (UZH, German Dept)

Anne Göhring (UZH, Institute CL)

Noemi Graf (UZH, German Dept)

Anja Hasse (UZH, German Dept)

Heath Gordon (UZH, CorpusLab)

Agnes Kolmer (UZH, URPP Language and Space)

Mike Lingg (UZH, German Dept)

Patrick Mächler (UZH, German Dept)

Josef Novak (Spitch)

Eva Peters (UZH, German Dept)

Uliana Petrunina (UZH, CorpusLab)

Hana Ruch (UZH, URPP Langauge and Space)

Beni Ruef (UZH, sms4science)

Tanja Samardžić (UZH, CorpusLab)

Yves Scherrer (LATL, Geneva)

Franziska Schmid (UZH, German Dept)

Fatima Stadler (UZH, CorpusLab)

Janine Richner-Steiner (UZH, German Dept)

Phillip Ströbel (UZH, CorpusLab)

Simone Ueberwasser (UZH, sms4science)

Alexandra Zoller (Spitch)