Lang Resources & Evaluation (2019) 53:735–769
https://doi.org/10.1007/s10579-019-09457-5

ORIGINAL PAPER

Digitising Swiss German: how to process and study a polycentric spoken language

Yves Scherrer[1] · **Tanja Samardžić**[2] · **Elvira Glaser**[3]

Published online: 11 April 2019

Abstract Swiss dialects of German are, unlike many dialects of other standardised languages, widely used in everyday communication. Despite this fact, automatic processing of Swiss German is still a considerable challenge due to the fact that it is mostly a spoken variety and that it is subject to considerable regional variation. This paper presents the ArchiMob corpus, a freely available general-purpose corpus of spoken Swiss German based on oral history interviews. The corpus is a result of a long design process, intensive manual work and specially adapted computational processing. We first present the modalities of access of the corpus for linguistic, historic and computational research. We then describe how the documents were transcribed, segmented and aligned with the sound source. This work involved a series of experiments that have led to automatically annotated normalisation and part-of-speech tagging layers. Finally, we present several case studies to motivate the use of the corpus for digital humanities in general and for dialectology in particular.

Keywords Swiss German · Corpus · Non-standard language · Spoken language · Normalisation · Speech-to-text alignment · Word level annotation · Dialectology · Dialectometry · Oral history

✉ Yves Scherrer
yves.scherrer@helsinki.fi

Tanja Samardžić
tanja.samardzic@uzh.ch

Elvira Glaser
eglaser@ds.uzh.ch

[1] Department of Digital Humanities, University of Helsinki, Helsinki, Finland

[2] Language and Space Lab, URPP Language and Space, University of Zurich, Zurich, Switzerland

[3] Department of German, University of Zurich, Zurich, Switzerland

1 Introduction

Swiss society is characterised by highly complex linguistic practices in comparison with other European countries. In addition to four official languages (German, French, Italian, Romansh, in the order of the number of speakers) and a wide range of other languages spoken by foreigners living in Switzerland (25% of the population, according to the Swiss Federal Statistical Office[1]), the local population speaks a great variety of local dialects. Unlike in other European countries, where dialect usage decreases in favour of standardised variants in most social domains, Swiss dialects are widely used in various domains, including education and public speech. This is especially true in the case of German dialects, which are the topic of this article.

Traditionally, the domains of use between standard German and Swiss German dialects have been divided according to the concept of *medial diglossia* (Kolde 1981; Siebenhaar and Wyler 1997), where standard German is used in written communication (and some institutionalised settings of oral communication) and the various dialects, fairly different from standard German, in spoken communication. Following this traditional division, Swiss German varieties are rarely studied outside of the narrowly focused research area of dialectology. With the development of computer-mediated communication, the traditional division between the two domains has become less clear, as Swiss varieties are increasingly written and recorded (Siebenhaar 2003). These developments call for and, at the same time, allow automatic processing of Swiss German for various purposes, including both research in digital humanities and developing practical applications.

In contrast to the increasing demand, basic tools for natural language processing of Swiss German texts are relatively undeveloped. Adapting existing tools developed for standard German has not proved successful. Explorative experiments (Hollenstein and Aepli 2014; Samardžić et al. 2015) have shown that even a small amount of data in Swiss varieties is more useful for training language processing tools than much larger data sets in standard German.

This paper presents an annotated corpus of spoken Swiss German, the ArchiMob corpus. We take advantage of natural language processing tools to provide additional annotation layers and show with some case studies that the corpus is suitable for research in language and humanities. We target specifically two issues: (a) the challenges of digitising a heterogeneous group of linguistic varieties that have no written tradition and (b) the opportunities that such a resource brings for the study of language and humanities in Switzerland.

Most of the existing Swiss German resources were developed in the context of dialectology and consist of isolated word types, such as the dialect lexicon *Idiotikon* (Staub et al. 1881) and the linguistic atlas of German-speaking Switzerland (Hotzenköcherle et al. 1962–1997; Christen et al. 2013); more recent digital resources in the same paradigm include Kolly and Leemann (2015). The

[1] https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/stand-entwicklung/alter-zivilstand-staatsangehoerigkeit.html.

Phonogrammarchiv of the University of Zurich[2] has a relatively rich collection of speech corpora, some of which range back more than 100 years. This archive is currently being processed in order to serve as a digital research resource, but this work is still in progress. The Bavarian Archive for Speech Signals provides speech corpora for "Regional varieties of German",[3] but its Swiss parts seem to include regionally accented High German rather than Swiss German dialect.[4]

In contrast, the resource presented here is one of the first multi-dialectal corpora available for Swiss German. As a corpus of continuous speech, it enables not only the analysis of formal linguistic features, but also of its content. Compared to two other recent corpora of Swiss German dialect—a corpus of SMS messages (Stark et al. 2009–2015) and a corpus of written texts (Hollenstein and Aepli 2014)—the ArchiMob corpus is larger, is aligned with the sound source and contains finer-grained metadata such as the dialect of the speaker. On the other hand, not all annotation layers of ArchiMob are verified manually. The ArchiMob corpus is also the only one that represents (transcribed) spoken language and features a particular content (historical narratives).[5]

This paper starts with a presentation of the content of the ArchiMob corpus and its modalities of access (Sect. 2). In Sect. 3, we describe the encoding and annotation layers of the corpus in more detail. Section 4 summarises our experiments of automating the annotation tasks. Finally, Sect. 5 presents six case studies that rely on the ArchiMob corpus to investigate various aspects of dialectal variation and variation in the content, showcasing the interest of such a resource for digital humanities in general.

2 The ArchiMob corpus: from oral history to a digital research resource

The original Archimob project was initiated by a filmmaker, Frédéric Gonseth, in 1998 and was conducted by the Archimob association.[6] The goal of this collaboration between historians and filmmakers was to gather testimonies of personal experiences of life in Switzerland in the period from 1939 to 1945. The resulting archive contains 555 recordings of interviews 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 regions of Switzerland and represent both genders,

[2] https://www.phonogrammarchiv.uzh.ch.

[3] https://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html.

[4] This assumption is based on the sample recording given on http://www.bas.uni-muenchen.de/forschung/Bas/BasRVG1eng.html.

[5] Once completed, the Phonogrammarchiv will provide a similar representation, potentially forming a data set for longitudinal studies.

[6] Archimob (archives de la mobilisation): http://www.archimob.ch/. We use the spelling *Archimob* for the association and the data collection project, and *ArchiMob* for the corpus.

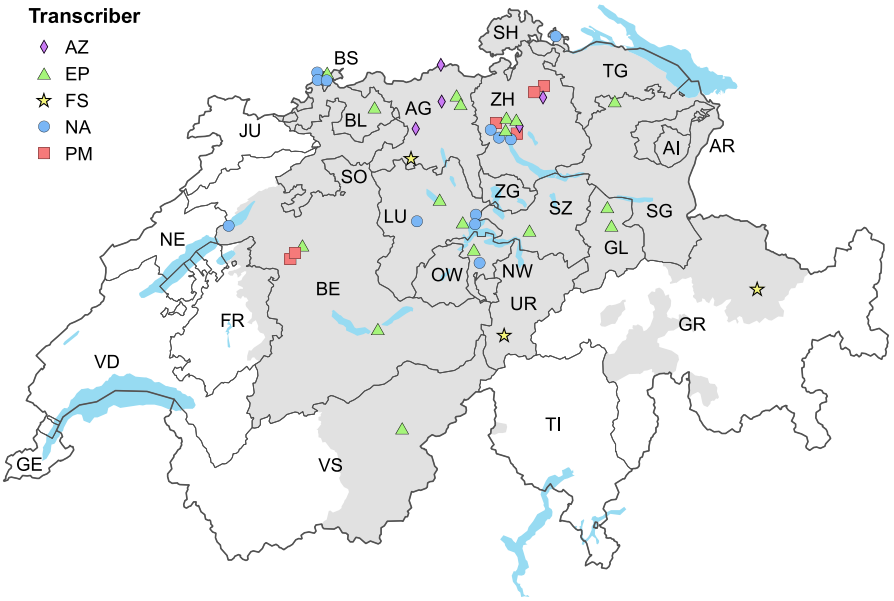

Fig. 1 Locations of the ArchiMob recordings included in the corpus, with different symbols for different transcribers (see Sect. 3.1). The grey area represents the German-speaking part of Switzerland. Canton boundaries are included as a proxy of major dialectal borders

different social backgrounds, and different political views. Most informants were born between 1910 and 1930.

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 goal of the corpus compilation was 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).

Of these 52 recordings, nine were excluded either because of poor sound quality or because the interviewees were highly exposed to dialect and language contact, making their productions less interesting for dialectological research. The remaining 43 recordings were then digitised into the MP4 format.[7]

The first release of the corpus contained 34 recordings transcribed with 15 540 tokens per recording on average (Samardžić et al. 2016). The second release, described in this article, contains all the 43 selected recordings, amounting to approximately 70 h of speech. Figure 1 shows the spatial distribution of the recordings included in the corpus, according to the origins of the speakers.

The selected recordings were then transcribed and processed so that they can be searched for diverse phenomena of interest to the researchers. The processing steps,

[7] The work described in this paper takes the digitised MP4 recordings as a starting point. While the digitisation to MP4 certainly has caused quality losses due to compression, we have not encountered any problems caused by the quality of the sound signal.

Fig. 2 An example of a query result with Sketch Engine for the word *gält* 'money'

described in more detail below, include writing normalisation and part-of-speech tagging.

To meet different needs of the users, we make the corpus accessible in two ways: as online look-up via corpus query engines and as an XML archive download. The current point of access to the corpus is its web page[8], but we consider integrating it in a larger infrastructure (such as CLARIN). The audio files are available on request.[9]

2.1 Online access with corpus query engines

After considering suitable corpus query engines for online look-up, we decided to use two systems, each with some advantages and disadvantages: Sketch Engine (Kilgarriff et al. 2014) and ANNIS (Krause and Zeldes 2014).[10]

An example of a search result with Sketch Engine is shown in Fig. 2. The system not only returns text passages with the exact match of the query word *gält* 'money', but also with dialectal variants such as *gäld* or *gäut*. Such a flexible search is made default in the simple search option. In order to relate different variants of the same word, we use normalised writing shown as grey subscript of the query word in Fig. 2. This normalised writing resembles standard German, but, as it is explained in more detail below (Sect. 3.2), it should not be considered an exact mapping between Swiss and standard German.

[8] http://www.spur.uzh.ch/en/departments/korpuslab/Research/ArchiMob.html.

[9] The XML archive contains some audio file samples.

[10] The corpus web page contains detailed information on how to access these systems.

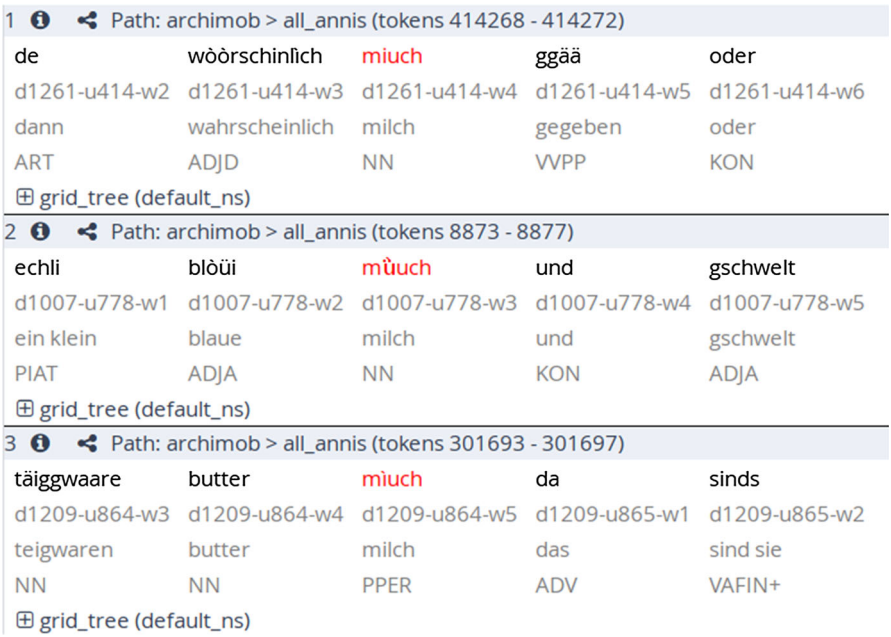

Fig. 3 An example of a query response with ANNIS for the normalised form *milch* 'milk', showing the dialectal variants *miuch* and *müuch*. Note that this example has been annotated automatically (see Sect. 4), illustrating some errors that might occur: in line 1, *de* should be tagged as ADV (adverb) instead of ART (article); in line 3, *mìuch* should be tagged as NN (noun) instead of PPER (personal pronoun). The normalised form *blaue* illustrates the difference between our normalisation and translation to standard German (correct translation would be *geschlagen*), described in more detail in Sect. 3.2

For a good functionality of our resource, it is crucial not to expect the user to know the exact normalisation of a word. The normalisation that we use is not a widely accepted standard in Switzerland and the user cannot be expected to know or to learn it.[11] To allow the users to search the corpus without knowing the normalisation, a new feature was implemented by the Sketch Engine team specifically for the purpose of our project. This feature allows the user to enter the query in any writing that seems plausible to her. If this writing occurred at least once in our corpus, we will be able to link and show instances of the queried item in all the other writings. Note that this approach to query is rather different from what used to be the practice in corpus query systems. Primarily conceived for working with text in standard languages, corpus query systems expect the user to know the exact writing for the query or to use regular expressions in order to approximate flexible search. Our solution enables searching resources with inconsistent writing in an intuitive and user-friendly way, making the resource accessible to a wider audience. This feature is thus potentially useful not only in the case of Swiss German but also for any non-standardised languages and varieties.

[11] As a matter of fact, the experience has shown that no single normalisation is likely to be widely accepted in Switzerland any time soon.

In addition to the new flexible search feature, Sketch Engine users can apply the standard functionality of the system to make more advanced queries using corpus query language, to manipulate resulting concordances, and to calculate different statistics (e.g., significant collocations).

To meet the need of some potential ArchiMob corpus users for a more detailed visualisation of search results, we use the ANNIS corpus query system. An example of an ANNIS query result is shown in Fig. 3. We can see that the system shows all the information currently available in the corpus at the same time. Below each transcribed word, we can see its corpus ID, normalised writing and part-of-speech tag. However, the number of shown hits needs to remain small as such a detailed view quickly fills up the screen.

We did not implement the flexible search option in ANNIS because this system is intended to be used by more advanced users interested in linguistic details. This purpose is reflected not only in the detailed responses of the system, but also in the rather advanced querying skills that are needed in order to perform any searches.

An important difference between the two corpus query systems is that users outside the European Union need to have a paid account in order to access our corpus with the Sketch Engine,[12] whereas personal accounts on ANNIS are free.

2.2 XML download

In addition to online look-up, we provide an XML archive for download. The XML format of the documents in the archive follows the Text Encoding Initiative (TEI) recommendations whenever possible. We add specific elements only for the cases not explicitly covered by TEI (e.g. the attribute `normalised`). This format is the base for producing the formats required by the corpus query engines. An overview of the steps performed in order to obtain the final XML format is given in Fig. 4. These steps are described in more detail in the following sections.

The data are stored in three types of files:

- *Content files* contain the text of the transcriptions.
- *Media files* contain the alignment between transcribed text and the corresponding audio files.
- *Speaker file* contains the socio-demographic information about the informants (region/dialect, age, gender, occupation) and the information about the speakers' roles in the conversation (interviewer, interviewee).

The content files are segmented into utterances. The references to the speaker and the media file are specified as attributes of each utterance (element "u"), as shown in the following illustration:

```
<u start="media_pointers#d1007-T176" xml:id="d1007-u88" who="person_db#EJos1007">
```

[12] A free-access, open-source version of the Sketch Engine is available under the name NoSketch (Rychlý 2007). For the moment, we do not use this version.

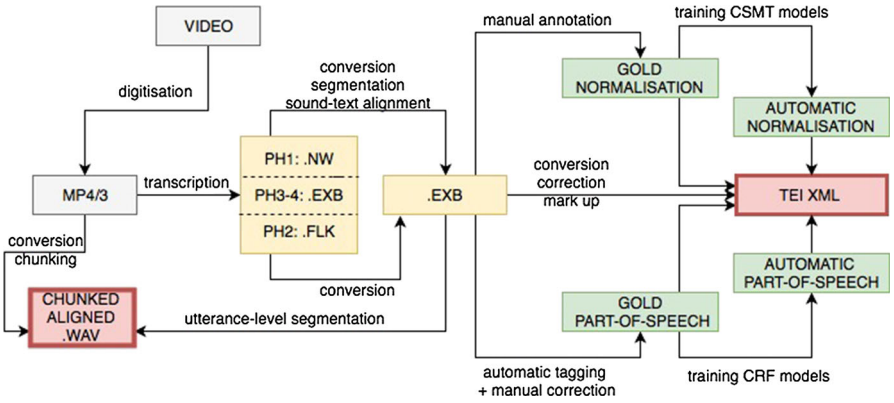

Fig. 4 Transcription and text processing work flow. Final (shared) data formats are marked with red background, intermediate text formats with beige, intermediate sound/video with grey, and annotation steps with green. (Color figure online)

Utterances consist of words (element "w"). Normalisation and part-of-speech tagging are encoded as attributes of the element "w", as in:

<w id="..." normalised="einest" POS="ADV" xml:id="..."> ainisch</w>

In addition to usual annotated words, utterances can contain pauses (vocalised or not), repeated speech, and unclear (or untranscribable) passages. Pauses are not counted as words; they are annotated with a different label (<pause xml:id="..."/>), as illustrated below. In repeated speech, the word in question is annotated as a word only once; the repeated fragments are annotated as deletion (<del> ... </del>). Unclear speech is annotated with a label that can span over multiple words.

<del type="truncation" xml:id="...">hundertvierz/</del>
<del type="truncation" xml:id="...">hundertvier/</del>
<w normalised="hundertfünfundvierzig" tag="NN" xml:id="...">hundertfiife-vierzgi</w>

The media and the speaker files are simple XML documents that consist of lists of time and speaker IDs respectively associated with the corresponding information.

3 Corpus encoding and annotation

Transforming oral history recordings into a widely accessible research resource requires extensive processing. In this section, we describe the encoding and annotation steps that were undertaken in creating the resource described above, underlining the challenges specific to Swiss German.

3.1 Transcription and speech-to-text alignment

The 43 documents selected for inclusion in the corpus contain approximately 70 h of speech. They were transcribed in four phases by five transcribers, with an average of 30 person-hours invested in transcribing 1 h of recordings. The modes and the phases of transcription were not part of a single plan, but rather a result of different circumstances in which the work on the corpus took place. Table 1 sums up the time line of the annotation process.[13]

The transcribed text is divided into utterances that correspond to transcription units of an approximate average length of 4-8 seconds and aligned to sound at this level. The utterances are mostly fragments spanning over one or more sentence constituents. We do not mark sentence boundaries. As it is usual in spoken language corpora, utterances are grouped into turns. We do not mark the boundaries between turns explicitly. Instead, we annotate utterances with speaker IDs. A change in the speaker (and its role) signals a turn boundary.

The transcription units, aligned with the sound source, are manually formed by transcribers. Such alignment is part of the output of specialised tools like FOLKER and EXMARaLDA (Schmidt 2012, for both tools). Since no specialised tool was used in phase 1, the 16 documents produced in this phase needed to be aligned subsequently. We approach this task by first automatically aligning the transcriptions with the sound source at the level of words using the tool WebMAUS (Kisler et al. 2012). To obtain the utterance level alignment comparable to the output of the transcription tools, we join the WebMAUS alignment automatically into larger units and then import it into EXMARaLDA for manual correction. For around one third of the transcriptions, the automatic alignment did not work well enough to be used as a pre-processing step. In these cases, we first produced an approximation of the target segments automatically based on the pauses encoded in the transcription. We then imported the transcriptions into EXMARaLDA for manual correction.

There is no widespread convention for writing Swiss German. We use the writing system "Schwyzertütschi Dialäktschrift" proposed by Dieth (1986), as is standard in recent dialectological research. The transcription is expected to show the main phonetic properties of the variety but in a way that is legible for everybody who is familiar with standard German spelling (Dieth 1986, 10). The function of the grapheme inventory in the Dieth's script depends on the dialect and its phonetic properties. 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 a considerable degree—also to the dialectal background of the transcriber.

Dieth's system, which is originally phonemic, can be implemented in different ways depending on how differentiated the phonetic qualities are to be expressed. The practice in using Dieth's system changed over the transcription phases, so that more distinctions concerning the openness of vowels were made in the first phase than in the later phases [e.g., phase 1: *èèr* vs. phase 3: *er* (std. *er*, engl. 'he')].

[13] Further information about the annotation guidelines and the geographic distribution of the documents can be found on the project web page.

Table 1 Overview of the four transcription phases

Phase	Years	Transcriber	Document IDs	Transcription tool
1	2006–2012	EP	1007, 1048, 1063, 1073, 1075, 1142, 1143, 1147, 1170, 1195, 1198, 1207, 1209, 1212, 1261, 1270	Nisus writer
2	2012–2014	PM	1082, 1083, 1087, 1121, 1215, 1225, 1244	FOLKER
3	2015	NA	1008, 1055, 1138, 1188, 1189, 1205	EXMARaLDA
		AZ	1228, 1248, 1259, 1295, 1300	
4	2016–2017	NA	1044, 1053, 1203, 1224, 1235, 1263	EXMARaLDA
		FS	1163, 1240, 1255	

A manual inspection of the transcriptions showed that these changes in the guidelines were not the only source of inconsistency. Different transcribers tended to make different decisions on how to implement the guidelines, not only regarding vowel quality, but also regarding word segmentation and other issues. These inconsistencies are one of the reasons why we introduce an additional annotation layer of normalised word tokens.

3.2 Normalisation

Variation in written Swiss German is observed at two levels. First, dialectal variation causes lexical units to be pronounced, and therefore also written, in a different way in different regions. Second, a lexical unit that can be considered phonetically invariant (within a region) is written in a different way on different occasions, due to occasional intra-speaker variation and, as mentioned above, to transcriber-related variation. In order to establish lexical identity of all writing variants that can be identified as "the same word"—needed to enable flexible search for instance—they need to be normalised to a single form.

Table 2 illustrates the range of potential variation with an arbitrarily chosen segment from our corpus. The table shows all the variants of the chosen words found in the same document, that is within a sample of the size of around 10 000 tokens transcribed by the same trained expert. In addition to these, more variants are found in other documents containing samples from other varieties. The shown variants include cases of regional variation (e.g., *gsait, gsä*it, gseit), variants due to changing transcription guidelines (e.g., *hed, hè*d) and variants caused by code-switching (e.g., *mä*in, main, main, mann, hat).

In the example of Table 2, all normalised forms correspond to standard German forms. Indeed, whenever a Swiss German word form corresponds to a standard German form in meaning and etymology, the standard German form is used for normalisation. However, it is important to note that we do not conceive normalisation as translation into standard German. A real translation to standard German would require substantial syntactic transformations and lexical replacement, whereas our normalisation is a word-by-word annotation of lexical identity

Table 2 A segment of a transcribed (original) text with corresponding variants found in the same document and in other documents

Original	*min*	*maa*	*het*	*immer*	*gsaait*
Variants in the same document	*mi*		*hat**		*gsait*
	mii				
	miin				
Variants in other documents (not exhaustive)	*mine*	*ma*	*hed*	*ime*	*gsäit*
	mìì	*man**	*hèd*	*imer*	*gsääit*
	*mäin**	*mann**	*hèt*	*emmer*	*gseit*
	*main**		*hät*	*imme*	*gseid*
	*mein**		*hätt*	*immers*	*ggsait*
Normalisation	*mein*	*mann*	*hat*	*immer*	*gesagt*
English	my	husband	has	always	said

Variants marked with * represent code-switched or cited standard German words (in contexts such as *mein Gott* 'my God', *Mein Kampf* 'my struggle', *Not am Mann* 'there is need', *Thomas Mann*). The common normalisation of all variants is shown in the Normalisation row

whose exact forms can be seen as arbitrary. Some of our choices for the normalisation language require further explanation:

- Swiss German word forms that do not have etymologically related standard German counterparts are normalised using a reconstructed common Swiss German form. For example, *öpper* 'someone' is normalised as *etwer* instead of the semantic standard German equivalent *jemand*, *töff* 'motorbike' to *tö*ff instead of standard German *motorrad*, *gheie* 'to fall' to *geheien* instead of *fallen*. Likewise, Swiss German *vorig* 'remaining' is normalised as *vorig*, even though this word means 'previous' in standard German.
- Standard German conventions regarding word boundaries are often not applicable to Swiss German, where articles and pronouns tend to be cliticised. As a result, transcribers often produce single tokens that correspond to several standard German tokens (albeit with a lot of transcriber-related variation). In such cases, we allow several tokens on the normalised side. For example, *hettemers* is normalised as *hätten wir es*, and *bimene* is normalised as *bei einem*.
- Sometimes, normalisation has the welcome side effect of disambiguating homophonous dialect forms. For example, *de* is normalised as *der* (definite article) or as *dann* (temporal adverb), depending on the context.
- In other cases, a normalised form encompasses formally distinct dialect forms, due to morphosyntactic syncretism. For example, the first normalised word of Table 2, *mein*, will also be applied to dialect forms such as *mis, miis*, which are neuter forms of masculine *min, miin*.

An important feature of our approach is that we regard normalisation as a hidden annotation layer used only for automatic processing. As discussed above, 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.

We describe our approach to normalisation in detailed guidelines, which we then apply to manual annotation of six documents taken from the transcription phase 1 (document IDs 1007, 1048, 1063, 1143, 1198, 1270) by three expert annotators. For this task, we used annotation tools that allowed annotators to quickly look up previous normalisations (if they exist) for the current word. We initially used VARD 2 (Baron and Rayson 2008), but we later switched to the better adapted SGT tool (Ruef and Ueberwasser 2013). These manually normalised documents were then used as a training set for automatic normalisation with character-level machine translation discussed in detail in Sect. 4.2.

3.3 Part-of-speech tagging

Annotation of part-of-speech tags is important for enabling more abstract queries in the corpus, regarding word classes and their combinations rather than concrete words. Part-of-speech tagging is a well studied task in NLP, and a rich offer of tools is available. These tools, however, are developed with written standardised languages in mind, while we need to apply them to a spoken non-standard variety.

We approach this task following Hollenstein and Aepli (2014), who adapted the widely used Stuttgart–Tübingen–Tagset (STTS) (Thielen et al. 1999) to a written version of Swiss German. The adaptation of the tag set addresses the following specific phenomena observed in Swiss German dialects:

- A new label PTKINF is introduced for the infinitival particles *go, cho, la, afa*. These particles are used when the respective full verbs (to go, to come, to let, to begin) subcategorise an infinitival clause. As this phenomenon does not exist in standard German, the addition of a new label is warranted.
- The label APPRART, used in standard German for preposition + definite article, is extended to preposition + indefinite article, as in *bimene* ‘at a’, which does not exist in standard German.
- The labels VAFIN+ and VMFIN+ apply to verb forms with enclitics. The latter usually are pronouns, e.g., *häts* ‘has it’, *hettemers* ‘would have we it’. Conjunctions with enclitics are labelled as KOUS+, e.g., *wemmer* ‘when we’.
- Whenever the *zu*-particle (phonologically reduced to *z* in Swiss German) is attached to the infinitive, the PTKZU+ tag is used: *zflüge* ‘to fly’.
- Adverbs with enclitics (which can be articles or other adverbs) are given the ADV+ tag: *sones* ‘such a’.

We apply this adapted tag set in manual annotation of three test sets:

- Test_0: 791 randomly selected segments from the same documents that are manually normalised (approximately 10%)
- Test_1: 600 randomly selected segments transcribed in the phases 1 and 2
- Test_2: 300 randomly selected segments transcribed in the phases 3 and 4

These test sets are used to assess the performance of different tagging models and strategies. By making separate tests, we intend to track potential effects of the variability in the corpus. Test_0 is the closest to the training data, which consist of the remaining 90% segments of the six documents included in manual normalisation, as described in more detail in Sect. 3.2. Test_1 and Test_2 are progressively more distant: Test_1 is partially transcribed by the same person as the documents used for training, while Test_2 contains the newest transcriptions.

4 Adaptation and evaluation of automatic processing tools

Our resource is intended to offer accurate and reliable information about the use of Swiss German. The size of the data set, however, should be big enough to allow quantitative analyses. We therefore aim to achieve the quality of manual encoding and annotation, but also to automate the processing steps to allow scaling up the data size. In Sect. 3, we described the encoding and annotation steps required to build the corpus. In this section, we describe the experiments with automatic systems adapted to perform these tasks.

4.1 Automatic transcription with automatic speech recognition

All the transcriptions included in the current version of the corpus are produced manually using the transcription tools listed in Table 1. However, these transcriptions now constitute an initial training set for future automatic processing with automatic speech recognition (ASR) systems.

To obtain a baseline for future improvements, we perform learning experiments with our current data set and a prototype speech recognition system developed in collaboration with a private company, an adaptation of the open-source Kaldi toolkit (Povey et al. 2011). We report here the results of this first evaluation.

The initial version of our ASR system is trained on the transcriptions representing the varieties of the larger Zurich area (see document IDs in Table 3). We choose this region as the largest relatively homogeneous subset of data containing around 48 h of speech. We start the training with a homogeneous sample in order to be able to assess the effects of adding heterogeneous samples to the training set at a later stage.

Table 3 Initial ASR evaluation

Train	Test	Precision	Recall	F-score
1007, 1055, 1063, 1082, 1083, 1138, 1143, 1147, 1188, 1189, 1195, 1198, 1205, 1207, 1209, 1228, 1244, 1248, 1259, 1270, 1295, 1300 (Larger Zurich area)	1170 (Bern)	48.57	22.35	29.83
	1263 (Basel)	61.16	36.11	45.40
	1240 (Grisons)	72.56	37.84	49.68
	1261 (Lucerne)	65.48	13.60	22.37
	1255 (Uri)	45.64	22.96	30.46
	1212 (Valais)	52.79	28.49	36.96

We test the system on six documents from different regions (see Table 3). For each document, we create 3-min samples starting: (a) in the middle of the document, (b) in the middle of the first half, and (c) in the middle of the second half. For each of the 18 samples, we manually count the following:

- *Gold transcription T*: the number of word tokens in the manual transcription of the sample.
- *System output O*: the number of word tokens (different from 'unknown') in the system output for the given sample.
- *Strict overlap S*: the number of word tokens that are identical in the system output and the manual transcription.
- *Flexible overlap F*: the number of word tokens in the system output that are not identical to the gold tokens, but still judged as correct by native speakers.

We then define the measures of precision and recall in terms of the collected counts. Precision is expressed as the proportion of correct word tokens, including the flexible overlap, in the system output ($\frac{S+F}{O}$). Recall is the proportion of words correctly recognised by the system in the gold transcription ($\frac{S+F}{T}$). F-measure is then calculated in the standard way. We take the average score over the three samples as a performance measure at document level. These average values are reported in Table 3.

Our evaluation shows that our ASR system is rather conservative: precision is systematically higher than recall. The variation in the performance over different documents (representing different regions) is considerable, but it is not explained solely by the known regional variation (discussed in more detail in the following section). In a subjective assessment by our transcribers, the current output of the system is judged not sufficient as a pre-processing for manual transcription.

We will therefore continue to work on improving the ASR by introducing more data and new learning methods. Improvements are possible for both the acoustic model and the language model. The acoustic model will benefit from a better representation of the phonetic features, while the language model will benefit from new training techniques including character-level modelling and neural networks. The baseline and the evaluation scheme described here will be crucial for monitoring improvements in the future.

4.2 Automatic normalisation with character-level statistical machine translation

The task of normalisation described in Sect. 3.2 is a recurring issue in dealing with different kinds of non-standard texts such as historical, spoken or computer-mediated communication (Dipper et al. 2013a, b; Bartz et al. 2013, e.g., for different non-standard varieties of German). Automatic word normalisation has been a popular topic in historical NLP over the last few years, resulting in a range of methods that are primarily useful for treating small edits in largely similar words (Baron and Rayson 2008; Bollmann 2012; Pettersson et al. 2013a).

More recently, character-level statistical machine translation (CSMT) has been successfully applied to normalisation of computer-mediated communication (De Clercq et al. 2013; Ljubešić et al. 2014) and historical texts (Pettersson et al. 2013b, 2014; Scherrer and Erjavec 2016). This method has originally been proposed for translation between closely related languages (Vilar et al. 2007; Tiedemann 2009). It requires less training data than word-level SMT but is limited to applications where regular changes occur at character level.

As for Swiss German dialects, word normalisation has already been manually performed by Stark et al. (2009–2015) using a collaborative annotation platform (Ruef and Ueberwasser 2013).

Our approach includes both manual and automatic annotation. We first normalise a small set of documents manually and train an automatic normalisation tool on these documents. After the initial evaluation, we try to improve the quality of the annotation in two ways. First, we explore improvements in the automatic methods. Second, we increase the training set by correcting manually some of the automatic output. With the iterative technical and manual improvement, we provide a good quality annotation that can be scaled up to larger data sets.

We choose CSMT for the automatic annotation because the string transformations that need to be performed in our case exceed the power of rule based or string-similarity methods. An alternative approach would be to use neural sequence-to-sequence methods (Cho et al. 2014; Sutskever et al. 2014). Neural methods are shown to outperform traditional statistical machine translation. However, experiments have not shown a clear advantage on the task of normalisation. While the recent shared task on normalisation of historical Dutch (Tjong Kim Sang et al. 2017) suggest that CSMT still performs better on this task, Honnet et al. (2017) have obtained better performance with neural methods. We intend to introduce neural methods in the future, examining these findings and exploring recent models especially suited for character-level string transformations in low input-data settings. These methods have been tested on morphological transformation tasks (Aharoni and Goldberg 2017; Makarov et al. 2017), but they can be extended to our normalisation task.

Table 4 shows the main steps in our adaptation of the standard CSMT for the task of dialect normalisation. In all the tests shown in the table, we use the system Moses (Koehn et al. 2007) with GIZA++ for word alignment.[14] We adapt the input for character-level calculations instead of the standard word level and we experiment with different system parameters and data sets.

We start by testing the default system settings using manually normalised documents (as discussed in Sect. 3.2). We first work with unchanged manual normalisation, which we term *pre-release* in Table 4. To account for the strong generalisation tendency of CSMT, we combine the output of CSMT system with simple memory-based learning. We take as the final output the most frequent normalisation for the test items seen in the training set and the CSMT output for the unseen items. This yields an accuracy score of 77.28% (Samardžić et al. 2015).

[14] GIZA++ is an implementation of the IBM alignment models (Brown et al. 1993).

Table 4 Step-by-step improvements in automatic normalisation with CSMT

Data/evaluation	Method	Accuracy (%)
Pre-release training data	Memory-based learning	77.28
1007, 1048, 1063, 1143, 1198, 1270	Word-by-word CSMT, 1 LM	
5-fold cross-validation	(Samardžić et al. 2015)	
Release training data	Memory-based learning	84.13
1007, 1048, 1063, 1143, 1198, 1270	Word-by-word CSMT, 1 LM	
5-fold cross-validation	(Samardžić et al. 2016)	
Release training data	Tuned segment-level CSMT, 2 LMs	90.46
1007, 1048, 1063, 1143, 1198, 1270	Constraints for memory-based learning	
10% held-out for testing	(Scherrer and Ljubešić 2016)	
Release training data + 1142, 1212	Tuned segment-level CSMT, 2 LMs	89.90
10% held-out for testing	No constraints	

During manual inspection of the first results, it turned out that the initial normalisation guidelines were not explicit enough to guarantee consistent annotation by the three annotators. For example, the unambiguous Swiss German form *dra* was sometimes normalised as *dran* and sometimes as *daran*; both normalisations are correct Standard German words. Also, the Swiss German form *gschaffet* was sometimes normalised to the semantic Standard German equivalent *gearbeitet* and sometimes to its etymological equivalent *geschafft*. We thus revised both the manual normalisation and the corresponding guidelines. For the examples cited above, we gave preference to the longer form *daran* and to *geschafft*. Furthermore, descriptions of non-vocalised communicative phenomena that had accidentally ended up as tokens in three of the texts were excluded from the normalisation task.

We reran the experiments with the same settings as above, but with the improved data set (termed *release* in Table 4 because this version is included in the official corpus release), and obtained a rise in the accuracy from 77.28 to 84.13%. Detailed results are reported by Samardžić et al. (2016). The observed improvement in accuracy underlines the importance of clear and easy-to-follow guidelines, especially for smaller datasets like ours.

We further improved the normalisation tool by a) tuning the CSMT system to optimise the weights for the translation model and for the language model, b) increasing the translation unit from a word to an entire utterance (the condition termed *segment* in Table 4), c) augmenting the training set for the language model with a corpus of spoken standard German (condition *LM2* in Table 4). These experiments, described in detail by Scherrer and Ljubešić (2016), lead to a considerable improvement in the performance cancelling the need for combining CSMT with memory-based learning. However, the best accuracy score of 90.46% is achieved only after introducing a constraint that selects the single observed normalisation for those test items that are seen in the training set with exactly one normalisation.

Fig. 5 Dialectal origin of the texts used in the CSMT experiments. The six initial training texts are displayed with red circles, whereas the two additional training texts are displayed with blue stars. (Color figure online)

To assess whether adding more Swiss German examples is beneficial to CSMT, we correct manually the automatic output in two documents and then add the corrected documents to the training and tuning data (the bottom row in Table 4). Figure 5 shows the locations of the six initial and the two added documents. Note that the benefits of adding more data are not evident in the context of a highly varied data set such as Swiss German (Samardžić et al. 2016). Nevertheless, we obtain comparable performance as in the previous setting without using additional constraints.

4.3 Part-of-speech tagging with adapted taggers and active learning

Part-of-speech annotation is in principle portable across similar languages (Yarowsky et al. 2001). This fact, together with the fact that similar tagged corpora already exist, brought us to the decision to start part-of-speech tagging of the ArchiMob corpus by adapting the existing models and tools.

There are two potential similar sources that could be used for training an initial part-of-speech tagging model, both with some advantages and disadvantages.

1. *TüBa-D/S* (Hinrichs et al. 2000) is a corpus of spontaneous dialogues conducted in standard German (360 000 tokens in 38 000 utterances). This corpus is of the same genre as ArchiMob (spoken language), but it is a different language variety (standard German vs. Swiss German).

Table 5 Results of the part-of-speech tagging experiments in terms of accuracy and out-of-vocabulary words (OOV)

Training	Test	% Accuracy	% OOV
TüBa-D/S	Normalised	70.68	24.21
NOAH's corpus	Original	73.09	30.72

2. *NOAH's* Corpus of Swiss German Dialects (Hollenstein and Aepli 2014) represents approximately the same variety (Swiss German), but it is small in size (73 000 tokens), from various sources of written language, and not normalised.

To assess which source provides better models for our ArchiMob data, we test them both on the Test_0 set (described in Sect. 3.3). We select test items from the documents that are manually normalised so that we can measure the performance on both original transcriptions and normalised words. Original transcriptions are closer to NOAH's corpus, while normalised writing is closer to TüBa-D/S. Both corpora contain punctuation, whereas the ArchiMob corpus does not. We therefore removed all punctuation signs for the purpose of our experiments.

Table 5 shows the main outcome of these initial evaluation experiments (more detailed results are reported by Samardžić et al. (2016)). Both results are obtained using the BTagger (Gesmundo and Samardžić 2012), which has shown good performance on smaller training sets.

Due to the relatively large training set and surface form similarity, we expected to obtain the best initial score by training a tagger on TüBa-D/S and testing on the normalised version of ArchiMob. This expectation, however, proved wrong, as we obtained a considerably better score by training on NOAH's corpus, despite the fact that this setting included much larger variation in writing (no normalisation is used) and that NOAH's corpus is relatively small. Although the proportion of test words unseen in training is larger in NOAH's setting (OOV in Table 5), the performance of the tagger is better.

We note that the additional tags introduced in NOAH's corpus to account for morphosyntactic particularities of Swiss German dialects (see Sect. 3.3) help produce better results. Indeed, 2.45% of tokens in the gold standard are tagged with one of the additional tags; the NOAH's tagger provides 68.05% accuracy on these tokens, whereas the TüBa-D/S tagger, having not seen the correct tags in the training data, gets them all wrong.

Following these findings, we set out to annotate more data in Swiss German by gradually adding manually corrected output of automatic tagging to the train set. Table 6 shows the steps in this process: we tag (still with BTagger) one document at the time, then correct it and add it to the train set in the next iteration. In every iteration, we note down the proportion of correctly tagged tokens in the new file before correction. We can see that the proportion of correct tags generally increases more in the first two than in the last two iterations.

Table 6 The increase of the part-of-speech tagging performance through correction of entire documents

Training	Test	% Correct
NOAH's	1007	77.18
NOAH's + 1007	1048	82.28
NOAH's + 1007, 1048	1063	87.32
NOAH's + 1007, 1048, 1063	1198	88.99
NOAH's + 1007, 1048, 1063, 1198	1270	92.51

With five ArchiMob documents added to the training set, we move on to improving the performance of the tagger. At this point, we replace BTagger with a conditional random fields (CRF) tagger, available as a Python library.[15] We decided to change the algorithm because the CRF tagger is a newer, better supported algorithm, more flexible, easier to use and embed in new tagging frameworks. A comparison of the performance of the two taggers showed that this change does not lead to a loss in the quality of the output.

We proceed with the improvements in two ways. First, we enrich the tagging model adding normalised forms, now available in the added ArchiMob documents, as features. Second, we increase the training data set by adding segments corrected through an active learning procedure.

The normalised forms used to enrich the model are annotated manually in the initial set of six documents described in Sect. 3.2 and in the extended set mentioned in Sect. 4.2. In other documents, we use the output of the automatic annotation (Sect. 4.2).

The active learning interactive annotator, developed for the purpose of this project by the TakeLab, University of Zagreb, runs the best current model on all currently unannotated utterances and identifies those items where the tagger is least confident. These items are presented to the human annotator for correction and then added to the training set for the next iteration. The procedure is repeated as long as it yields improvements on the test sets.

The active learning component is developed to meet the specific needs of our project: the user is presented with a pre-tagged low-confidence utterance and is asked to correct the tags that are wrong. Since these units are rather short, the interface displays a number of previous utterances, in order to provide enough context for the user to evaluate the tags with longer dependencies. The previous utterances are presented as simple text with no annotation.

The interface is run from the command line. It is configurable by means of an accompanying Python script, where the user can set: (a) the number of segments to annotate in one iteration, (b) the length of the previous context, (c) the span for the tagger's hyper-parameter optimisation. These settings are made configurable because they depend on the size of the existing training set and on the time available for training the tagger and entering new annotations. As the annotation advances, the settings need to be adapted to ensure optimal use of the interface. The interface also allows separating the task of training the model from the annotation

[15] https://python-crfsuite.readthedocs.io.

Table 7 Accuracy scores (%) obtained in part-of-speech tagging experiments with CRF and ArchiMob data only

	Test_0	Test_1	Test_2
Plain CRF	84.4	76.4	74.2
+ normalisation	92.3	85.4	79.9
+ normalisation + AL_1 (100)	92.6	86.0	80.6
+ normalisation + AL_2 (300)	92.5	86.0	83.8

task. In this way, we can run the two components according to our own time schedule and goals.

The evaluation outcomes for the automatic part-of-speech tagging with the described improvements are shown in Table 7.[16] These experiments show that using the normalisation feature helps the tagger even in the cases of Test_1 and Test_2, where this annotation is noisy (automatic output without manual correction). As for the increase of the training data with active learning, we observe most benefits in the case of Test_2, where the performance is lower than on the other two sets. We can also see that the improvements are proportional to the number of corrected items (100 segments in the third row in Table 7 vs. 300 segments in the last row). As the performance approaches the threshold of 90% accuracy score, the impact of training data increase becomes limited.

5 Studying linguistic variation using the ArchiMob corpus

The ArchiMob corpus is not only an interesting object of study for computational linguistics, it can also serve as a precious resource for dialectological and historical research, as has been intended from the beginning of the project. In this section, we present several case studies to illustrate the potential of the ArchiMob corpus. In Sect. 5.1, we investigate to what extent dialectal variation can be captured by looking at the transcriptions alone. Section 5.2 asks similar questions about linguistic variation, but tries to answer them by taking into account the normalisations. Section 5.3 illustrates how the annotations can be used to investigate the content of the texts.

For all case studies, we only use the utterances produced by the informants, not those produced by the interviewers. By doing so, we hope that the data material is as representative of the informant's dialect as possible. Also, we remove diacritics from the phase 1 transcriptions in order to control for the most obvious effect of gradual changes to the guidelines.

5.1 Extracting dialectal variation patterns from speech transcriptions

The transcriptions of the ArchiMob corpus provide an interesting dataset for detecting dialectal variation patterns. Here, we discuss the tasks of identifying the

[16] All the results shown in the table are obtained using the ArchiMob data only. We do not show the performance obtained with the data from the NOAH's corpus in the training set because they were generally inferior to those with the ArchiMob data only.

dialectal origin of an utterance (dialect identification, Sect. 5.1.1) and of classifying the documents according to their linguistic similarity (dialect classification, Sect. 5.1.2).

5.1.1 Dialect identification

Language identification is an important task for natural language processing in general. While relatively simple methods perform well for languages that are sufficiently different, language identification for closely related languages is still a challenging task (e.g., Zampieri et al. 2014). Identifying the origin of dialect texts can be viewed as a particular case of this problem. In this spirit, data from the ArchiMob corpus were used to set up the *German Dialect Identification* task at the VarDial 2017 and 2018 workshops (Zampieri et al. 2017, 2018).

Four dialectal areas with a sufficient number of texts and which were known to be distinct enough were used in the identification tasks: Zurich (ZH), Basel (BS), Bern (BE), and Lucerne (LU). For each dialect area, utterances from at least three documents were selected as training data, and utterances from a different document were chosen for testing the systems.

Ten teams participated in the 2017 task, and eight teams in the 2018 task. Most participants obtained between 60% and 70% macro-averaged F1-scores. These figures are probably not far away from human performance: some utterances do not contain any dialect-specific cues and therefore cannot be reliably classified even by experts. It remains to be seen to what extent the addition of acoustic data can make up for lacking detail in the transcriptions. Further details about the task setup, the submitted systems and the obtained results can be found in the respective publications (Zampieri et al. 2017, 2018).

5.1.2 Dialect classification

Inspired by the dialect identification task, we wanted to extend this idea to the more general problem of automatic dialect classification, by (a) taking into account all documents of the corpus, and (b) not relying on predefined dialect areas. Concretely, we wanted to investigate to what extent dialect areas could be inferred directly from the data.

Uncovering dialect areas is one of the main goals of dialectometry (e.g., Goebl 1982, 1993). The traditional dialectometrical pipeline consists of the following steps (after Goebl 2010, 439):

- The linguistic data, typically extracted from a dialectological atlas, is formatted into a *data matrix* of n enquiry points $\times$ m linguistic features. Each cell contains the local variant of a feature at an enquiry point.
- A *distance matrix* of n points $\times$ n points is derived from the data matrix, by pairwise comparison of the feature vectors of two enquiry points. The distance matrix typically is symmetric, with 0 values on the diagonal.

- Then, a dimensionality reduction algorithm is applied to reduce each row of the distance matrix to a single value (or a small number of values v), leading to a *value matrix* of n points $\times$ v values. A wide variety of algorithms have been proposed, and one of the simplest ones (also used in the following) is hierarchical clustering, which assigns each enquiry point a cluster ID, grouping the points with the most similar linguistic characteristics together.
- The values of the value matrix are colour-coded and plotted onto a map. The hypothesis is that geographically close places will be clustered together, and where they are not, a dialectological explanation for this mismatch will have to be found.

The dialectometrical pipeline has mainly been applied to atlas data (Goebl 2005), where each column in the data matrix contains a feature known to vary across dialects. Introduction of corpus data into the study of regional variation allows collecting information about text frequency of the varying forms and constructions as they are spontaneously produced (Wolk and Szmrecsanyi 2016). The main disadvantages of this data source are uneven spatial coverage (naturally occurring texts tend to be more concentrated in particular regions) and sparseness of linguistic phenomena (the features of interest typically show only rarely in text). For instance, only 114 normalised word types are realised in all documents of the ArchiMob corpus. In addition to this, variation across recordings of free conversations, such as ArchiMob, may be due to personal preference and context, and not necessarily to dialectal variation. Studying linguistic distance using corpus data therefore requires departing from a typical dialect data matrix.

In a pilot study (Scherrer 2012), we explicitly tried to match words with similar transcriptions across dialects, using a preliminary version of the ArchiMob corpus. Here, we propose to use language modelling, a technique that has also been used in the dialect identification task (Gamallo et al. 2017b), to create a distance matrix directly. The last steps of the dialectometrical pipeline can then be applied as before.

In particular, we create a language model for each document of the ArchiMob corpus and show how well it fits all other documents of the collection. The assumption is that a language model will better fit a text of the same dialect than a text of a distant dialect. We estimate character 4-gram language models using the KenLM tool (Heafield 2011) with discount fallback, and we use document-level perplexity as a distance measure (Gamallo et al. 2017a). The resulting "distance" matrix is not symmetrical, as the perplexity of model A on text B is not guaranteed to be identical to the perplexity of model B on text A. Likewise, the diagonal does not necessarily contain 0 values, as the perplexity of model A on text A is not always equal to 0. For classification, we apply hierarchical clustering with Ward's algorithm.[17]

[17] A range of other visualization methods from different areas of digital humanities may be applied here, e.g. language interaction networks as in Gamallo et al. (2017a), force-directed graph layouts as used in phylogenetics (Jäger 2012), bootstrap consensus trees as in the stylometric study of Rybicki and Heydel (2013), induced decision trees (Gibbon 2016), or various alternative methods used in dialectometry (for an overview, see e.g. Wieling and Nerbonne 2015).

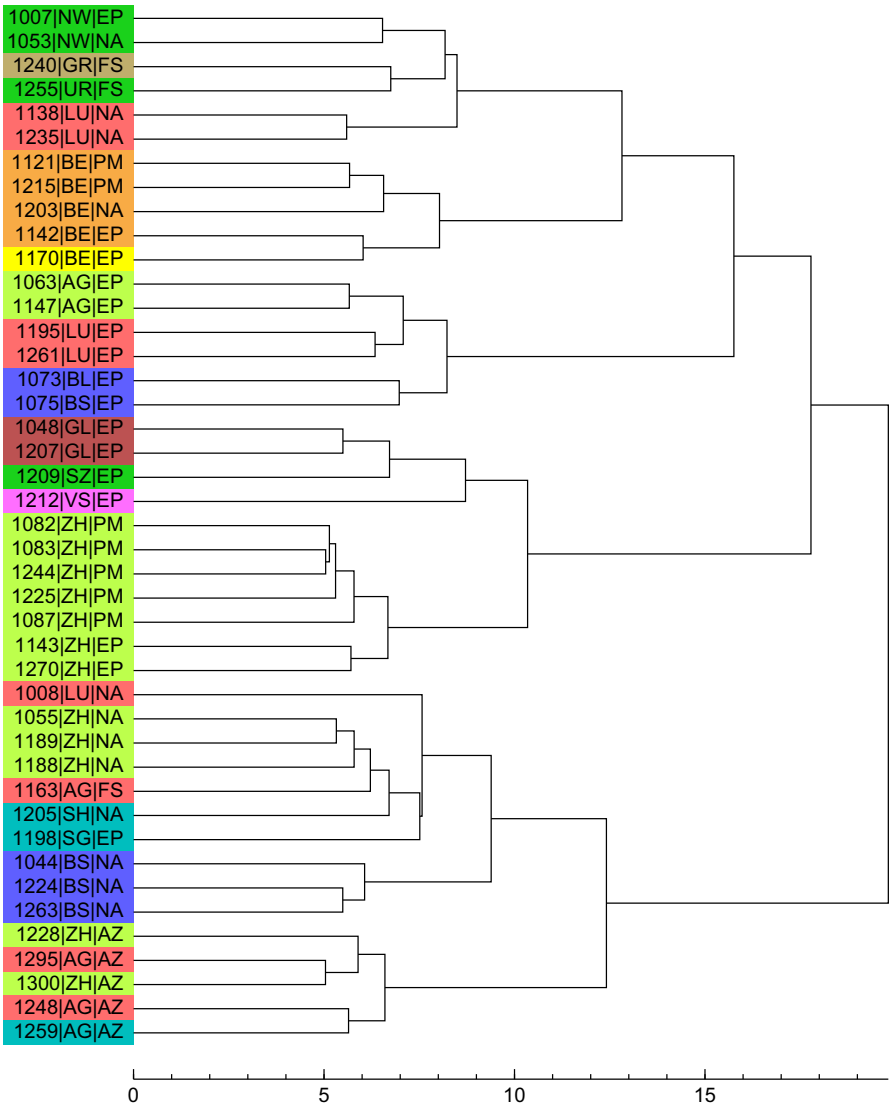

Fig. 6 Dendrogram of the Ward clustering of language model perplexities. Each leaf is labelled with the document ID, the canton of its origin, and the transcriber initials. The background colors refer to the dialect regions inferred by Scherrer and Stoeckle (2016). (Color figure online)

Figure 6 shows the results of the clustering in the form of a dendrogram. It can be seen that the transcriber effect is quite strong: documents edited by the same transcriber tend to cluster together. The texts from the Zurich area (ZH) are partitioned into three distinct areas, according to the transcriber. Nevertheless, some

dialectologically interesting groupings can be found: BL and BS refer to similar dialects, NW and UR as well, and VS is least connected to any other dialect area.

The clustering obtained with the ArchiMob documents can be compared with a similar experiment based on data from two Swiss German dialect atlases according to the traditional dialectometrical approach (Scherrer and Stoeckle 2016): each ArchiMob document in Fig. 6 is assigned a color, which corresponds to the cluster inferred at that geographical location by (Scherrer and Stoeckle 2016). The matching suggests that at least a subset of documents contains a dialectologically differentiated signal that can rival with much more cost-intensive atlas data.

As transcriber effects cannot be eliminated completely (partly also because of the transcription guideline changes), future work will focus on the statistical modelling of transcriber variation and dialectal variation as distinct effects (Wieling et al. 2011).

5.2 Normalisation as basis for dialectological comparison

In the previous section, we referred to the difficulty of comparing features in the different dialect texts: not all speakers use the same linguistic structures, but it is difficult to tease apart proper dialectological effects from subject-induced and personal preferences. However, the normalisation layer can help here: all linguistic elements (words, graphemes or characters) that are normalised the same way can be compared with each other. In this section, we explore two applications of this idea, one related to particular (phonological) phenomena, and one related to aggregate dialect measurements.

5.2.1 Investigating phonological variation in dialect texts

Investigating phonological properties in transcribed speech is challenging, and even more so if the transcription guidelines are known to have changed over time and transcriber differences are known to be prominent. Despite these challenges, we show in the following that known phonological variation patterns can be efficiently searched and compared across documents thanks to the normalisation layer.

Taking several methodological shortcuts, we define a phonological variable as a grapheme on the normalisation layer, and its possible dialectal realisations (variants) as the set of graphemes on the transcription layer it is aligned with. For example, the phonological variable represented by the normalisation grapheme *ck* has two levels, the dialectal variants *k* and *gg*, whose frequency distribution varies according to the origin of the texts.[18] For this definition to work, we a) need to align characters between transcription tokens and normalisation tokens, and b) group adjacent characters into multi-character graphemes when required.

A popular character alignment technique is based on Levenshtein distance, where the edit operations that contribute to the Levenshtein distance calculation are

[18] According to the Dieth spelling guidelines, the grapheme ⟨*k*⟩ reflects the pronunciation [kx], whereas the grapheme ⟨*gg*⟩ reflects [k:]. The phonetic realisation of the normalisation graphemes is not relevant here.

converted to alignment links. Several extensions have been proposed for dialect data, e.g., by prohibiting alignments of vowels with consonants (Wieling et al. 2009).

Character-level statistical machine translation (CSMT), which we already used for normalisation, is an alternative to these approaches. It does not presuppose any notion of word (or character) identity and thus works equally well with different character inventories or writing systems. The most widely used alignment models were proposed in the early days of statistical machine translation (Brown et al. 1993) and have been used for character alignment in our CSMT setting. Since character alignments are an integral part of the CSMT translation models, we can extract from our normalisation models any alignments of interest. We thus align characters and extract grapheme correspondences using exactly the same process as for creating the CSMT normalisation models, except that we create a distinct model for each document.

Graphemes do not always consist of single characters. Character sequences that frequently co-occur and that are frequently aligned in the same way should be grouped together as a multi-character grapheme. This process has also been studied in the field of statistical machine translation under the name of phrase extraction (Och et al. 1999), and can again be straightforwardly converted from the word level to the character level. The *phrase table* file created during CSMT model training lists all grapheme pairs together with their (co-)occurrence counts, allowing us to easily compute relative frequencies of the transcription graphemes.

Let us return to the example given above and examine how the normalised grapheme *ck* is realised in two arbitrarily chosen ArchiMob documents:

- Document 1: k 37.0%, gg 63.0%
- Document 2: k 95.2%, gg 2.4%, ch 2.4%

This analysis can be extended to all texts of the ArchiMob corpus, and the frequency distributions of each variant can be plotted on a map. Figure 7 shows such a plot for the *gg* variant. The frequencies extracted from the ArchiMob texts can be compared with atlas data from the *Sprachatlas der deutschen Schweiz* (SDS; Hotzenköcherle et al. 1962–1997). The area of use of the *gg* variant according to the atlas (map 2/095) is reproduced as a green background on Fig. 7. Seven ArchiMob documents show relative frequencies higher than 0.5 for the *gg* variant. All these documents are located in the three regions where the atlas data also shows the *gg* variant: Basel (Northwest), St. Gallen (Northeast), and Glarus (Southeast).

Another interesting phenomenon is the vocalisation of intervocalic *ll* in Western Swiss German dialects. Figure 8 shows the relative frequencies of the vocalic variant *u* in the ArchiMob texts, and the occurrence of the same variant according to atlas data. Again, one can see that all ArchiMob speakers who vocalise are located in (or near) the areas where the atlas predicts vocalisation. However, the frequency values are spread widely. In the case of the three texts of the Bern area, this variation reflects—at least to some extent—the sociolinguistic status of *l*-vocalisation as a lower-class phenomenon (Siebenhaar 2000): the lower-class

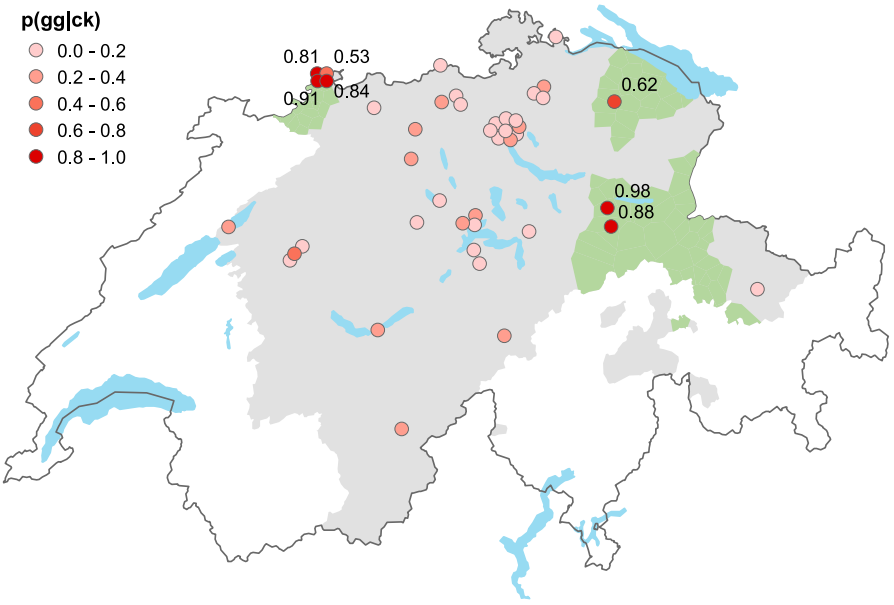

Fig. 7 Probabilities of *ck* dialectally realised as *gg*. The green areas represent the distribution of the *gg* variant in SDS map 2/095 *drü*cken 'to push'. (Color figure online)

speaker, a gardener, uses vocalisation more often (69%) than the two upper middle-class speakers, a draughtsman (33%) and a doctor (49%). In Central Switzerland, two documents exhibit vocalisation, and both stem from the edges of the vocalisation area as defined by the SDS map. There are also some ArchiMob documents from within the vocalisation areas that do not show any evidence of this phenomenon. Whether this mismatch should be attributed to language change or to annotation effects remains to be analysed.

The two examples given above show that the geographic extension of some phonological variants extracted from the ArchiMob corpus coincide remarkably well with those of the Swiss German dialect atlas SDS. As the ArchiMob interviewees are about one generation younger than the informants of the SDS, the proposed technique can also be used to trace dialect change. For example, Christen (2001) has found *l*-vocalisation to extend eastwards to the city of Lucerne and Nidwald (the Southern shore of Lake Lucerne), but the ArchiMob documents from that area do not show any vocalisation. This suggests that this linguistic change may have set in more recently.

However, the method presented here is limited by the precision of the transcription: obviously, only variation patterns that are reflected in the transcription can be retrieved. For example, studies on the realisation of */r/* cannot be carried out (at least not without analysing the corresponding audio data) as the different variants are not distinguished in the transcription. Likewise, studies on vowel quality will not be reliable as not all documents of the ArchiMob corpus are transcribed in the same way. Still, we believe that the proposed approach can shed a new light on dialect

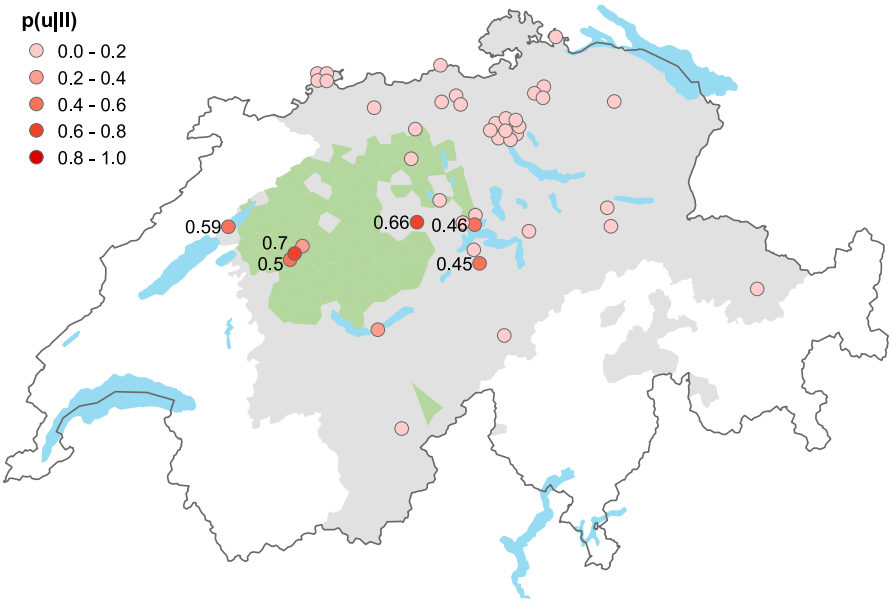

Fig. 8 Probabilities of *ll* dialectally realised as *u*. The green areas represent the distribution of the *u* variant in SDS map 2/198 'Teller'. (Color figure online)

variation and change in German-speaking Switzerland, complementing detailed phonetic information that can be found in other resources.

5.2.2 Dialectality measurements

The normalisation layer can also be used to perform aggregate analyses of the ArchiMob data. The case study presented here is inspired by a measure known as *dialectality*, a score that expresses the distance between a dialect text and the standard variety (Herrgen and Schmidt 1989; Herrgen et al. 2001). This method has seen a lot of success in Germany, where small-scale dialects are in the process of being replaced by larger-scale regiolects. Dialectality has proved to be a relevant measure of the degree of advancement of this process.

The dialectality measure requires character-aligned phonetically transcribed data in the dialect and the standard language. The phonemes are compared pairwise, and for each phoneme pair a distance value is computed, based on the number of phonetic features that need to be changed. These distance values are then averaged across words and utterances to provide a single dialectality value for each text.

We simplify this idea drastically for our purposes. First, we assume the normalisation layer to be our standard language, which is not quite accurate. Second, we do not attempt to convert the transcriptions and normalisations into true phonetic transcriptions, as they are generally underspecified. Instead, we use plain Levenshtein distance to compute the distance value per word. Figure 9 plots the dialectality values of all ArchiMob texts.

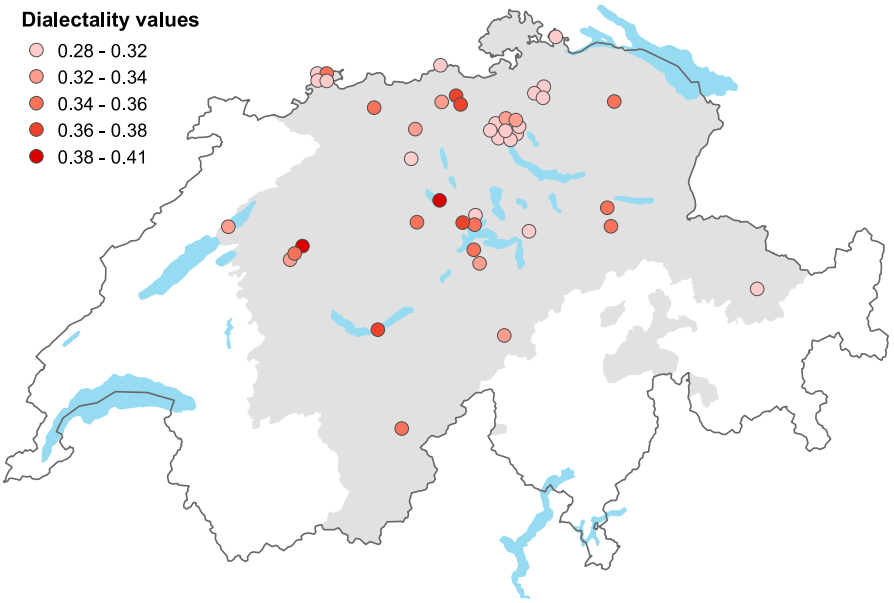

Fig. 9 Dialectality values. (Color figure online)

The results show that the dialectality values do not differ much between documents. Also, they seem neither strongly correlated with geography nor with the transcriber. With the exception of the northernmost documents located very close to the German border, it shows that Northern dialects are not subject to higher assimilation pressure from standard German than southern Swiss dialects are. This is expected, as the standard variety is perceived in Switzerland as a vertical counterpart in the diglossia, not a "horizontal" counterpart on the geographical (in our case) North-South axis (Siebenhaar and Wyler 1997).

The lowest dialectality values tend to be found in the Zurich area, which suggests that the Zurich dialect acts as a sort of default dialect with a low number of characteristic traits. This effect has been found in several studies based on different datasets (Scherrer and Rambow 2010; Hollenstein and Aepli 2015; Scherrer and Stoeckle 2016).

5.3 Content analysis

In the previous sections, we have focused on the form of the linguistic data in the ArchiMob texts, leading naturally to interpretations in the field of dialectology. Another type of analysis, with potentially much larger impact in the field of digital humanities, refers to the content of the texts. Qualitative analyses could be carried out using methodologies from conversation analysis and interactional sociolinguistics. Quantitative analyses can also be envisaged, and two illustrations of the potential of the dataset for such quantitative analyses are given in the following.

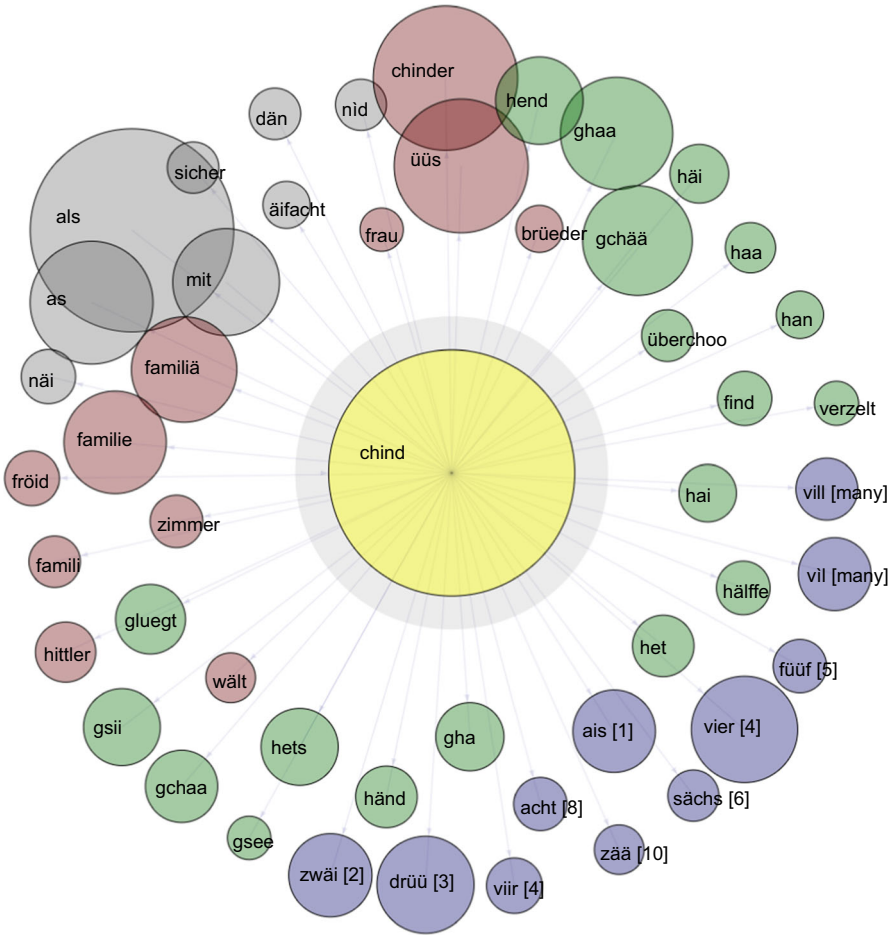

Fig. 10 Collocates of the word *chind* in the ArchiMob corpus. (Color figure online)

5.3.1 Collocation analysis

Collocation analysis (or more generally co-occurrence analysis) is a simple and popular method in digital humanities for measuring associations between entities and concepts in texts. For each keyword, the most frequently co-occurring concepts are found, using weighting techniques from information retrieval. Collocation analysis can be performed on a whole corpus, or separately for different partitions of the corpus, in order to find changes in usage patterns.

The SketchEngine tool provides a built-in collocation analysis method, which can be used interactively by account holders. As an example, Fig. 10 shows a collocation analysis for the word *chind* 'child/children',[19] performed over all

[19] The singular and plural forms are homophonous in most Swiss German dialects.

documents of the ArchiMob corpus. The collocates are visualised in a circle around the keyword, with different colours representing different parts-of-speech. By looking at numerals (purple circles) alone, the ArchiMob corpus can provide us with an interesting insight into the traditional family sizes in Switzerland in the first half of the 20th century. Words relating to other concepts and their associations can be analysed in a similar way.

5.3.2 Lexical change

Schifferle (2017) convincingly demonstrates the usefulness of a corpus such as ArchiMob for lexicological studies. Changes in the usage of words, in turn, hint at societal changes and can therefore be potentially interesting for a wide range of disciplines such as sociology, psychology or political science.

In his research, Schifferle investigates the usage of the relationship terms *koleeg* 'colleague, friend' and *früend* 'friend' in the 16 ArchiMob texts of phase 1. He finds that *koleeg* is used nearly exclusively by male speakers, whereas *früend* is used almost exclusively by female speakers. This is partly an effect of the traditional gender role distribution where men were more likely to have work and military colleagues than women, but it also hints at some kind of taboo for male speakers regarding the use of *früend* (or the feminine term *früendin*, also included in the study) for non-romantic relationships.

A second result concerns the contexts of use of *koleeg*. While recent dictionaries specify that Swiss German usage is not restricted to 'work/club/military colleague' but can refer to a close personal friendship, there is no historical evidence of this usage in the relevant lexicons. Although this usage is only marginally attested in the ArchiMob sample examined by Schifferle, exactly these occurrences may provide a precious insight into the emergence of this semantic shift.

6 Summary of contributions

In this paper, we introduce a general-purpose research resource consisting of manual transcriptions of oral history recordings in Swiss German, the ArchiMob corpus. We present its construction and functionality and show examples of its use for a range of research topics in digital humanities.

We have retraced the history of the construction of this corpus. The discontinuous work on this corpus, under different responsibilities, using different (sometimes not well adapted) tools, resulted in various types of inconsistencies, which we tried to minimize through various correction and harmonisation rounds. We have also presented additional annotation layers like normalisation and part-of-speech tagging, for which we were able to obtain competitive automatic annotation results in a difficult setting, emphasizing the usefulness of language technology for digital humanities.

A recurrent characteristics of this corpus—as alluded to in several applications described in Sect. 5—is transcriber inconsistency. Consistency of transcriptions is

indeed a central point in dealing with dialect corpora in particular and with dialectological data in general (Mathussek 2016). We take this issue into account carefully. The large amount of variation in transcription is a central reason for providing a normalisation level of annotation; any research question that does not explicitly rely on the phonological realisation of the words and utterances can be addressed on the basis of the normalisation level alone. This layer allows more systematic studies of variance in writing as different versions of the same word can be identified and analysed.

Furthermore, the XML documents are annotated with the transcriber identification, which allows the researcher to create subcorpora that are minimally affected by this issue. While considerable effort has been spent on harmonizing the transcriptions since the first release of the corpus (Samardžić et al. 2016),[20] we sketch the potential application of automatic speech recognition (Sect. 4.1) to create virtual transcriptions that can then be compared with the existing manual ones. Future work will show if this approach leads to significant improvements. While the transcriber effects do influence some findings on the regional variability (Sect. 5), we address these effects directly by comparing our classifications with those performed using atlas data.

To summarize, we argue that a resource such as the ArchiMob corpus is an important reference for new quantitative approaches to the study of language use and variation, not only in dialectology, but also in social sciences and history. Our solutions to the challenges of encoding and annotating a polycentric spoken language constitute a collection of know-how that can facilitate further developments of similar resources. Finally, the tools for automatic processing that we adapted and evaluated are now available for use and further improvements in future development of similar resources, but also in more general processing of Swiss German recordings and texts.

Acknowledgements We would like to thank our numerous collaborators who participated in the development of this corpus: Noëmi Aepli, Henning Beywl, Christof Bless, Alexandra Bünzli, Matthias Friedli, Anne Göhring, Noemi Graf, Anja Hasse, Gordon Heath, Agnes Kolmer, Mike Lingg, Patrick Mächler, Eva Peters, Beni Ruef, Hanna Ruch, Franziska Schmid, Larissa Schmidt, Fatima Stadler, Janine Steiner-Richter, Phillip Ströbel, Simone Ueberwasser, Alexandra Zoller. Funding was provided by Hasler Stiftung (Grant No. 16038). Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

[20] Between release 1 and the present release, 0.03% of tokens (1200 out of 37,250) have been modified due to guideline violations. Many more tokens may be affected by some type of transcriber inconsistency, but these are difficult to spot without explicitly retranscribing the data.

References

Aharoni, R., & Goldberg, Y. (2017). Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics* (Vol. 1: long papers, pp. 2004–2015). Vancouver: Association for Computational Linguistics.

Baron, A., & 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*.

Bartz, T., Beißwenger, M., & Storrer, A. (2013). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL*, *28*(1), 157–198.

Bollmann, M. (2012). (Semi-)automatic normalization of historical texts using distance measures and the Norma tool. In *Proceedings of the second workshop on annotation of corpora for research in the humanities (ACRH-2), Lisbon, Portugal* (pp. 3–14).

Brown, P. E., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, eighth workshop on syntax, semantics and structure in statistical translation, Doha, Qatar, 25 October 2014* (pp. 103–111).

Christen, H. (2001). Ein Dialektmarker auf Erfolgskurs: Die /l/-Vokalisierung in der deutschsprachigen Schweiz. *Zeitschrift für Dialektologie und Linguistik*, *1*(68), 16–26.

Christen, H., Glaser, E., & Friedli, M. (2013). *Kleiner Sprachatlas der deutschen Schweiz*. Frauenfeld: Huber.

De Clercq, O., Desmet, B., Schulz, S., Lefever, E., & Hoste, V. (2013). Normalization of Dutch user-generated content. In *Proceedings of RANLP 2013, Hissar, Bulgaria* (pp. 179–188).

Dieth, E. (1986). *Schwyzertütschi Dialäktschrift*, 2nd edn. Sauerländer, Aarau, edited by Christian Schmid-Cadalbert.

Dipper, S., Donhauser, K., Klein, T., Linde, S., Müller, S., & Wegera, K. P. (2013a). HiTS: ein Tagset für historische Sprachstufen des Deutschen. *JLCL*, *28*(1), 85–137.

Dipper, S., Lüdeling, A., & Reznicek, M. (2013b). NoSta-D: A corpus of German non-standard varieties. In M. Zampieri & S. Diwersy (Eds.), *Non-standard data sources in corpus-based research, no. 5 in ZSM-Studien, Shaker* (pp. 69–76).

Friedli, M. (2012). Der Komparativanschluss im Schweizerdeutschen: Arealität, Variation und Wandel. Ph.D. thesis, Universität Zürich. https://doi.org/10.5167/uzh-68746.

Gamallo, P., Pichel, J. R., & Alegria, I. (2017a). From language identification to language distance. *Physica A: Statistical Mechanics and Its Applications*, *484*, 152–162.

Gamallo, P., Pichel, J. R., & Alegria, I. (2017b). A perplexity-based method for similar languages discrimination. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)* (pp. 109–114). Valencia: Association for Computational Linguistics.

Gesmundo, A., & Samardžić, T. (2012). Lemmatisation as a tagging task. In *Proceedings of the 50th annual meeting of the association for computational linguistics* (Vol. 2: short papers, pp. 368–372). Jeju Island: Association for Computational Linguistics.

Gibbon, D. (2016). Legacy language atlas data mining: Mapping Kru languages. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Paris: European Language Resources Association (ELRA).

Goebl, H. (1982). *Dialektometrie. Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

Goebl, H. (1993). Dialectometry: A short overview of the principles and practice of quantitive classification of linguistic atlas data. In R. Köhler & B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 277–315). Dordrecht: Kluwer.

Goebl, H. (2005). Dialektometrie (art. 37). In R. Köhler, G. Altmann, & R. G. Piotrowski (Eds.), *Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch/An International Handbook, Handbücher zur Sprach- und Kommunikationswissenschaft 27* (pp. 498–531). Berlin: De Gruyter.

Goebl, H. (2010). Dialectometry and quantitative mapping. In A. Lameli, R. Kehrein, & S. Rabanus (Eds.), *Language and Space. An International Handbook of Linguistic Variation, Handbücher zur Sprach- und Kommunikationswissenschaft 30.2* (Vol. 2, pp. 433–457). Berlin: De Gruyter.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the EMNLP 2011 sixth workshop on statistical machine translation* (pp. 187–197). Scotland: Edinburgh.

Herrgen, J., & Schmidt, J. E. (1989). Dialektalitätsareale und dialektabbau. In W. Putschke, W. H. Veith, & P. Wiesinger (Eds.), *Dialektgeographie und Dialektologie. Günter Bellmann zum 60. Geburtstag von seinen Schülern und Freunden, Deutsche Dialektgeographie 90* (pp. 304–346). Marburg: Elwert.

Herrgen, J., Lameli, A., Rabanus, S., & Schmidt, J. E. (2001). Dialektalität als phonetische Distanz. Ein Verfahren zur Messung standarddivergenter Sprechformen. http://archiv.ub.uni-marburg.de/es/2008/0007/pdf/dialektalitaetsmessung.pdf.

Hinrichs, E. W., Bartels, J., Kawata, Y., Kordoni, V., & Telljohann, H. (2000). The Tübingen treebanks for spoken German, English, and Japanese. In W. Wahlster (Ed.), *Verbmobil: Foundations of speech-to-speech translation* (pp. 550–574). Berlin: Springer.

Hollenstein, N., & 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), Association for Computational Linguistics, COLING 2014, Dublin, Ireland*.

Hollenstein, N., & Aepli, N. (2015). A resource for natural language processing of Swiss German dialects. In *Proceedings of GSCL, German society for computational linguistics and language technology* (pp. 108–109). Germany: Duisburg-Essen.

Honnet, P. E., Popescu-Belis, A., Musat, C., & Baeriswyl, M. (2017). Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. ArXiv e-prints arXiv:1710.11035.

Hotzenköcherle, R., Schläpfer, R., Trüb, R., & Zinsli, P. (Eds.). (1962–1997). *Sprachatlas der deutschen Schweiz*. Bern: Francke.

Jäger, G. (2012). Estimating and visualizing language similarities using weighted alignment and force-directed graph layout. In *Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH* (pp. 81–88). Avignon: Association for Computational Linguistics.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., et al. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36.

Kisler, T., Schiel, F., & Sloetjes, H. (2012). Signal processing via web services: The use case WebMAUS. In *Proceedings digital humanities 2012* (pp. 30–34). Hamburg, Germany.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 demonstration session, Prague, Czech Republic* (pp. 177–180).

Kolde, G. (1981). Sprachkontakte in gemischtsprachigen Städten. Vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i.Ue. No. 37 in Zeitschrift für Dialektologie und Linguistik, Beihefte, Steiner, Wiesbaden.

Kolly, M. J., & Leemann, A. (2015). Dialäkt Äpp: Communicating dialectology to the public—Crowdsourcing dialects from the public. In A. Leemann, M. J. Kolly, V. Dellwo, & S. Schmid (Eds.), *Trends in phonetics and phonology. Studies from German-speaking Europe* (pp. 271–285). Bern: Peter Lang.

Krause, T., & Zeldes, A. (2014). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, *31*, 118–139.

Ljubešić, N., Erjavec, T., & Fišer, D. (2014). Standardizing tweets with character-level machine translation. In *Proceedings of CICLing 2014, Springer, Kathmandu, Nepal. Lecture notes in computer science* (pp. 164–175).

Makarov, P., Ruzsics, T., & Clematide, S. (2017). Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 shared task: Universal morphological reinflection* (pp. 49–57). Vancouver: Association for Computational Linguistics.

Mathussek, A. (2016). On the problem of field worker isoglosses. In M. H. Côté, R. Knooihuizen, & J. Nerbonne (Eds.), *The future of dialects* (pp. 99–116). Berlin: Language Science Press.

Och, F. J., Tillmann, C., & Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT conference on empirical methods in natural language processing and very large corpora* (pp. 20–28).

Pettersson, E., Megyesi, B. B., & Nivre, J. (2013a). Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic conference of computational linguistics (Nodalida 2013)* (pp. 163—79). Norway, Oslo.

Pettersson, E., Megyesi, B. B., & Tiedemann, J. (2013b). An SMT approach to automatic annotation of historical text. In *Proceedings of the NoDaLiDa workshop on computational historical linguistics* (pp. 54–69). Oslo, Norway.

Pettersson, E., Megyesi, B. B., & Nivre, J. (2014). A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th workshop on language technology for cultural heritage, social sciences, and humanities (LaTeCH)* (pp. 32–41). Gothenburg, Sweden.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., et al. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

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. Ph.D. thesis, Universität Zürich. https://opac.nebis.ch/ediss/20121398.pdf.

Ruef, B., & Ueberwasser, S. (2013). The taming of a dialect: Interlinear glossing of Swiss German text messages. In M. Zampieri & S. Diwersy (Eds.), *Non-standard data sources in corpus-based research* (pp. 61–68). Aachen: Shaker Verlag.

Rybicki, J., & Heydel, M. (2013). The stylistics and stylometry of collaborative translation: Woolf's night and day in polish. *Literary and Linguistic Computing*, *28*(4), 708–717.

Rychlý, P. (2007). Manatee/Bonito—A modular corpus manager. In *1st workshop on recent advances in Slavonic natural language processing* (pp. 65–70). Brno: Masaryk University.

Samardžić, T., Scherrer, Y., & 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, ELRA* (pp. 294–298).

Samardžić, T., Scherrer, Y., & Glaser, E. (2016). ArchiMob—A corpus of spoken Swiss German. In N. C. C. Chair, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. Paris: European Language Resources Association (ELRA).

Scherrer, Y. (2012). Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH* (pp. 63–71). Avignon: Association for Computational Linguistics.

Scherrer, Y., & Erjavec, T. (2016). Modernising historical Slovene words. *Natural Language Engineering*, *22*(6), 881–905.

Scherrer, Y., & Ljubešić, N. (2016). Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th conference on natural language processing (KONVENS 2016)* (pp. 248–255).

Scherrer, Y., & Rambow, O. (2010). Word-based dialect identification with georeferenced rules. In *Proceedings of EMNLP* (pp. 1151–1161). Cambridge, MA: Association for Computational Linguistics.

Scherrer, Y., & Stoeckle, P. (2016). A quantitative approach to Swiss German-Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, *24*(1), 92–125.

Schifferle, H. P. (2017). Helvetische Beziehungen? Gschpäändli, Koleege, Fründ. Beziehungsbezeichnungen im Schweizerdeutschen. In A. Linke & J. Schröter (Eds.), *Sprache und Beziehung, Linguistik - Impulse & Tendenzen 69* (pp. 183–206). Berlin: De Gruyter.

Schmidt, T. (2012). EXMARaLDA and the FOLK tools. In *Proceedings of LREC 2012, ELRA, Istanbul*.

Siebenhaar, B. (2000). Stadtberndeutsch. In B. Siebenhaar & F. Stäheli (Eds.), *Stadtberndeutsch - Sprachschichten einst und jetzt, Schweizer Dialekte in Text und Ton 5.1* (pp. 7–32). Murten: Licorne Verlag.

Siebenhaar, B. (2003). Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats. Linguistik online 15. https://bop.unibe.ch/linguistik-online/issue/view/200.

Siebenhaar, B., & Wyler, A. (1997). *Dialekt und Hochsprache in der deutschsprachigen Schweiz* (5th ed.). Zürich: Pro Helvetia.

Stark, E., Ueberwasser, S., & Ruef, B. (2009–2015). *Swiss SMS corpus*. University of Zurich. https://sms.linguistik.uzh.ch.

Staub, F., Tobler, L., Bachmann, A., Gröger, O., Wanner, H., Dalcher, P., Ott, P., & Schifferle, H. P. (Eds.), (1881–) Schweizerisches Idiotikon: Wörterbuch der schweizerdeutschen Sprache. Huber, Frauenfeld 1881–2012/Schwabe, Basel 2015ff. http://www.idiotikon.ch.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014, December 8–13, 2014, Montreal, Quebec, Canada* (pp. 3104–3112).

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

Tiedemann, J. (2009). Character-based PSMT for closely related languages. In *Proceedings of EAMT 2009* (pp. 12–19). Barcelona, Spain.

Tjong Kim Sang, E., Bollmann, M., Boschker, R., Casacuberta, F., Dietz, F., Dipper, S., et al. (2017). The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation. *Computational Linguistics in the Netherlands Journal*, *7*, 53–64.

Vilar, D., Peter, J. T., & Ney, H. (2007). Can we translate letters? In *Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic* (pp. 33–39).

Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. *Annual Review of Linguistics*, *1*, 243–264.

Wieling, M., Prokić, J., & Nerbonne, J. (2009). Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 workshop on language technology and resources for cultural heritage, social sciences, humanities, and education* (pp. 26–34). Association for Computational Linguistics.

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, *6*(9), 1–14.

Wolk, C., & Szmrecsanyi, B. (2016). Top-down and bottom-up advances in corpus-based dialectometry. In M. H. Côté, R. Knooihuizen, & J. Nerbonne (Eds.), *The future of dialects: Selected papers from Methods in Dialectology XV, Language Variation 1*. Berlin: Language Science Press.

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on human language technology research* (pp. 1–8). Stroudsburg, PA: Association for Computational Linguistics.

Zampieri, M., Tan, L., Ljubešić, N., & Tiedemann, J. (2014). A report on the DSL shared task 2014. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects* (pp. 58–67). Dublin: Association for Computational Linguistics and Dublin City University.

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., & Aepli, N. (2017). Findings of the VarDial evaluation campaign 2017. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)* (pp. 1–15). Valencia: Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J., et al. (2018). Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the fifth workshop on NLP for similar languages, varieties and dialects* (pp. 1–17).