Proceedings Chapter

ArchiMob - A Corpus of Spoken Swiss German

SAMARDZIC, Tanja, SCHERRER, Yves, GLASER, Elvira

Abstract

Swiss dialects of German are, unlike most dialects of well 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 rarely recorded and that it is subject to considerable regional variation. This paper presents a freely available general-purpose corpus of spoken Swiss German suitable for linguistic research, but also for training automatic tools. The corpus is a result of a long design process, intensive manual work and specially adapted computational processing. We first describe how the documents were transcribed, segmented and aligned with the sound source, and how inconsistent transcriptions were unified through an additional normalisation layer. We then present a bootstrapping approach to automatic normalisation using different machine-translation-inspired methods. Furthermore, we evaluate the performance of part-of-speech taggers on our data and show how the same bootstrapping approach improves part-of-speech tagging by 10% over four rounds. Finally, we present the [...]

Reference

SAMARDZIC, Tanja, SCHERRER, Yves, GLASER, Elvira. ArchiMob - A Corpus of Spoken Swiss German. In: Calzolari, N. ; Choukri, K. ; Declerck, T. ; Goggi, S. ; Grobelnik, M. ; Maegaard, B. ; Mariani, J. ; Mazo, H. ; Moreno, A. ; Odijk, J. & Piperidis, S. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France : European Language Resources Association (ELRA), 2016.

ArchiMob — A corpus of Spoken Swiss German

Tanja Samardžić[1], **Yves Scherrer**[2], **Elvira Glaser**[3]
[1] URPP Language and Space, University of Zurich, Freiestrasse 16, 8032 Zurich,
[2] LATL-CUI, University of Geneva, Route de Drize 7, 1227 Carouge,
[3] German Department, University of Zurich, Schönberggasse 9, 8001 Zurich
tanja.samardzic@uzh.ch, yves.scherrer@unige.ch, eglaser@ds.uzh.ch

Abstract

Swiss dialects of German are, unlike most dialects of well 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 rarely recorded and that it is subject to considerable regional variation. This paper presents a freely available general-purpose corpus of spoken Swiss German suitable for linguistic research, but also for training automatic tools. The corpus is a result of a long design process, intensive manual work and specially adapted computational processing. We first describe how the documents were transcribed, segmented and aligned with the sound source, and how inconsistent transcriptions were unified through an additional normalisation layer. We then present a bootstrapping approach to automatic normalisation using different machine-translation-inspired methods. Furthermore, we evaluate the performance of part-of-speech taggers on our data and show how the same bootstrapping approach improves part-of-speech tagging by 10% over four rounds. Finally, we present the modalities of access of the corpus as well as the data format.

Keywords: Swiss German, corpus, non-standard language, spoken language, normalisation, speech-to-text alignment, word level annotation

1. Introduction

The term Swiss German covers a range of German varieties spoken in Switzerland on around two-thirds of its territory. Swiss German dialects are widely used in speech, while standard German is used nearly exclusively in written contexts.

This paper presents a general-purpose corpus of spoken Swiss German suitable for studying linguistic micro-variation and spatial diffusion with quantitative approaches, but also for developing natural language processing tools.

The compilation of this corpus is set in the context of increasing presence of Swiss German variants in different domains of everyday communication. This trend results in an accumulation of language materials (TV and radio recording, written blogs, personal communication through various popular channels) and in an increased interest in automatic processing.

As opposed to other, more or less digitised, sources of Swiss German data (Hotzenköcherle et al., 1962-1997; Staub et al., 1881- ; Scherrer and Rambow, 2010; Kolly and Leemann, 2015), which consist of isolated word types, this corpus is intended to represent continuous speech, that is, the words as they are actually used in texts. The main difference between the corpus presented in this paper and the other two existing corpora of Swiss German, a corpus of SMS messages (Stark et al., 2009-2015) and a corpus of written texts (Hollenstein and Aepli, 2014), is the fact that this is the only corpus of transcribed spoken language. Another corpus based on transcription of spoken language is under development; it contains a smaller sample of old recordings of shorter texts (more information is available at `http://www.phonogrammarchiv.uzh.ch/en.html`).

2. Data source and the size of the corpus

The corpus contains transcriptions of video recordings collected by the ArchiMob association (see `http://www.archimob.ch`) in the period 1999-2001. This collection contains 555 recordings. Each recording is produced with one informant using a semi-directive technique and is between 1h and 2h long. Informants come from all linguistic regions of Switzerland and represent both genders, different social backgrounds, and different political views.

In order to select the material to be transcribed for the ArchiMob corpus, the recordings were rated as category A, B, or C according to the linguistic representativeness of the speakers (speakers who were not exposed to dialect/language contact are considered most representative) and the sound quality. All the recordings of the category A, a total of 44, were selected to be included in the ArchiMob corpus. This release of the corpus contains 34 transcribed recordings, with 15 540 tokens per recording on average (around 500 000 tokens total).

3. Processing steps

Building a corpus of spoken language requires both intensive manual work and computational processing. In this section, we describe the used procedures and discuss the challenges specific to Swiss German.

3.1. Transcription and segmentation

The selected documents were transcribed in three phases by four transcribers. 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. In the first phase (2006-2012), 16 documents were transcribed without any specific tool. In the second phase (2011-2013), 7 documents were transcribed using FOLKER. The remaining 11 documents were transcribed

#373	schwöschter die isch sibe jar elter / aber / eh /	nüd	di gliich mueter / ìm sini mueter isch
#395	chliichind / drùm isch dän der altersunderschiid	nùùd	</u> <u> wa hend iri eltere gmacht </u>
#509	kategorii ine wänd iischwèngge </u> <u> mmh / nu	nöd	z früe (aber) </u> <u> aha </u> <u> (
#570	/ für nüüerige oder / eh / äifach er hät	nùd	/ am alte fescht gchläbet èèr hät äu gsee
#687	wider vù der gmäind verloosed woorde das	nùd	immer der gliich di gliiche gchaa hät /
#732	puure z reden und soo aber / eh / das hät me	nùd	vil anepraacht es bìzeli gröösser hät mes
#925	/ das isch dän ebe sini iischtellig gsìì	nùd	(4:57) / wisoo de daas was wit ez machche
#987	phunggt mönds üüs der erloo / das zalemer	nùùd	/ und das hät er dän anepraacht / (naher
#1038	esoo siini iischtellige gsìì für fortschrit	nöd	/ das er äifach / eh / nùd am alte äifach
#1046	für fortschrit nöd / das er äifach / eh /	nùd	am alte äifach nuur / chläbe pliben isch
#1055	am alte äifach nuur / chläbe pliben isch	nùd	</u> <u> wi grooss isch den die puurerai
#1093	überhäupt niemert esoo grooss we hüt / das isch	nùd	der fal gsìì / me hät öppe zwölf schtugg
#1205	äis guet das hät vilicht öppe / je wäiss	nùùd	öppe driize (?) heggtaar / hät daas händ
#1404	wider händ für über e winter / daas langet	nùùd	fürne scheesewage / so isch das gloffe
#1412	nùùd fürne scheesewage / so isch das gloffe	nöd	</u> <u> aso händ si s gfüül ghaa si sind
#1434	ufgwachse (?) das </u> <u> näi mir sind	nùd	aarm uufgwachsen aber mer hät ääifach für
#1444	uufgwachsen aber mer hät ääifach für öppis wo	nùd	üübedingt hät möse sii hät me möse verzichte
#1502	isch / aber ich / eh / ha ds gfüül das hät	nùd	gschat me hät eender d fantasii aagfangen
#1602	emal chù luege was du tüegisch / chasch ez	nùd	rede mitmer ich bi det ùnd det </u> <u>
#1654	lang / di riichscht gmäind gsìì òiroopa /	nùd	nu i de schwiiz / und dän dur d hüraaterii

Figure 1: A sample of corpus concordances; the query retrieves different words normalised as *nicht* 'not'.

in the third phase (2015) using EXMARaLDA (Schmidt, 2012, for both tools).

As there is no widely accepted orthographic standard for Swiss German, we use a set of recommendations proposed by Dieth (1986). However, the implementation of the recommendations changed over the time. More fine-grained phonetic distinctions are marked in the first transcriptions (e.g. a distinction between open and closed vowels), while the later ones are closer to orthographic transcriptions where only the most prominent contrasts are marked.

A manual inspection of the transcriptions showed that this was not the only source of inconsistency. However, we have decided not to change the transcriptions in order to make them more consistent. Instead, we add a level of annotation which allows us to establish the identity of variants, as discussed in more detail below.

The transcribed text is divided into utterances that correspond to transcription units of an approximate average length of 4-8 seconds. 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. Since no specialised tool was used in the first phase, the 16 documents produced in this phase needed to be aligned subsequently. We approach this task by first aligning automatically the transcriptions with the sound source at the level of words using the tool WebMAUS (Kisler et al., 2012). To obtain the level of 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.

3.2. Normalisation

The lack of written tradition in Swiss German causes considerable inconsistency in writing. The transcription recommendations by Dieth, although often used in expert transcriptions, tend to be interpreted and implemented in different ways, resulting in inconsistencies not only between the different transcription phases, but even within a single text transcribed by the same trained expert. Another source of inconsistency is considerable regional variation. Many words that are felt to be the same are pronounced and therefore written differently across regions. In order to establish lexical identities between the items felt like "the same word", the transcribed texts need to be normalised.

We approach this task in a semi-automatic fashion. We first normalise a small set of documents manually and train an automatic normalisation tool on these documents. In a second step, we use the automatic tool in a bootstrapping approach to pre-process additional documents.

3.2.1. Initial experiments

A set of six documents were normalised manually by three expert annotators. To ensure the consistency of annotation, we produced guidelines which listed case-based decisions. We also used annotation tools that allowed annotators to quickly look up previous normalisations for each word which had already been normalised. We initially used VARD 2 (Baron and Rayson, 2008), but we later switched to the better adapted SGT tool (Ruef and Ueberwasser, 2013). For more details on the approach to the manual and automatic normalisation, see Samardžić et al. (2015).

Then, an automatic normalisation tool was trained on these documents, using a combination of memory-based learning, character-level machine translation and language modelling (Samardžić et al., 2015). Evaluation using cross-validation yielded an average accuracy of 77.28%.

Training data	Initial		Initial		Initial		Initial		Augmented	
Test data	Cross-validation		Test1		Test2		Test3		Test3	
	Prop.	Acc.	Prop.	Acc.	Prop.	Acc.	Prop.	Acc.	Prop.	Acc.
Unique	46.69	98.13	33.63	96.50	43.02	98.60	41.82	95.28	43.50	95.33
Ambiguous 1	41.14	82.21	46.56	76.55	42.67	88.27	37.11	78.45	37.13	78.40
Ambiguous 2	0.49	61.14	0.25	52.38	0.72	69.79	0.38	75.56	0.52	77.42
New	11.68	35.90	19.57	50.40	13.59	51.47	20.70	44.45	18.85	42.01
All	100	84.13	100	78.08	100	87.58	100	78.44	100	78.90

Table 1: Results of the automatic normalisation experiments. The table shows, for all four considered word classes, the proportion of words per class (*Prop.*) and the normalisation accuracy of the class (*Acc.*). The *Initial* training set consists of 6 manually annotated documents. The *Augmented* training set contains the 6 initial documents plus the manually corrected *Test1* and *Test2* documents.

3.2.2. Normalisation correction

During manual inspection of the first results, it turned out that the normalisation guidelines were not explicit enough (or not followed thoroughly 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*.

In order to obtain normalisations that were as consistent as possible, we manually unified differring normalisations wherever this was desirable. For the examples cited above, we gave preference to the longer form *daran* as this was specified for similar cases in the guidelines, and to *geschafft* as the guidelines mark a preference for etymology over semantics. As stated above, these corrections were only applied to the normalisations, not to the transcriptions of the original forms. 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 normalisation experiments with the cleaned data and were able to raise the accuracy from 77.28% to 84.13% using the same methods. Detailed results are given in Table 1 (leftmost columns) according to the four word classes defined in (Samardžić et al., 2015): *Unique* words are associated with exactly one normalisation in the training set; *Ambiguous 1* words are associated with more than one normalisation candidate, but a unique most frequent normalisation can be determined; for *Ambiguous 2* words, no single most frequent normalisation can be selected because of tied frequency counts; and *New* words have not been observed during training and therefore no normalisation candidates are available.

The observed improvement in accuracy underlines the importance of clear and easy-to-follow guidelines, especially for smaller datasets like ours.

3.2.3. Bootstrapping

We use the automatic normalisation tool in a bootstrapping approach to pre-process additional documents, with the underlying assumption that it is faster to correct the errors of the automatic tool than to completely normalise the document by hand. Judging from the accuracy obtained in the cross-validation experiment, we assume that only 20 to 25% of tokens have to be corrected manually.

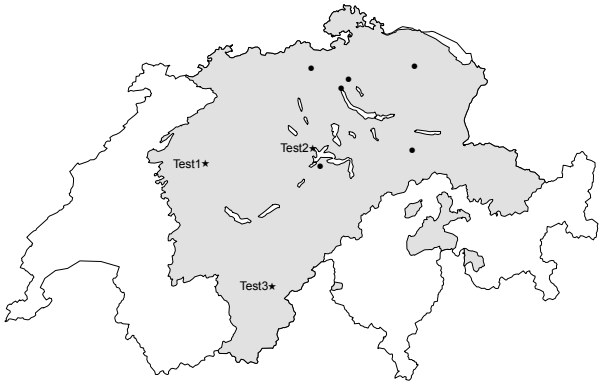

Figure 2: Dialectal origin of the texts used in the experiments. The six training texts are displayed with a circle, whereas the three test texts are displayed with a star. The German-speaking area of Switzerland is highlighted in grey.

In the first bootstrapping round, we automatically normalised two additional documents, *Test1* and *Test2* and then manually corrected them. Our assumption was true, since 78.08% and 87.58% of words in these two documents did not need to be manually corrected. The decrease in accuracy observed with *Test1* was expected, since this document comes from a dialect that has not been seen in the training, which also shows in the higher proportion of unknown (*New*) words in this text. In contrast, automatic normalisation showed even better accuracy for *Test2* than for cross-validation; this text was dialectologically more closely related to the ones used for training the tool. The dialectal origin of the texts is depicted in Figure 2.

In the second bootstrapping round, the normalisation tool is retrained by adding the manually corrected versions of *Test1* and *Test2* to the training data. An additional document, *Test3*, was automatically normalised using the initial as well as the augmented tool, and the latter version was then hand-corrected. Here, we wanted to test whether the retrained normalisation tool was able to improve the accuracy compared to the initial normalisation tool. One

effect of retraining is the increase of (easy to normalise) *Unique* candidates and the decrease of (difficult to normalise) *New* candidates, as detailed in the respective *Prop.* columns of Table 1. However, accuracy did not improve as much as expected, mainly because the character-level machine translation model responsible for the *New* translations did not perform as well as in the first round. To sum up, there is indeed a slight, although not statistically significant ($\chi^2(1; N = 23\,663) = 0.754; p = 0.385$). increase in accuracy, probably because that the dialect of *Test3* is not close enough to the dialects of *Test1* and *Test2*.

The bootstrapping process then continues by adding other corrected texts to the training set, retraining the normalisation tool, and applying it to the next documents. An important issue in choosing new texts for bootstrapping is their different dialectal origin. We plan to add texts in a principled way by gradually increasing the covered dialect area, starting with the (rather compact) area defined in the initial dataset. This work is currently in progress.

As soon as a sufficient number of normalised documents become available, it will be envisaged to build dialect-aware normalisation models. Each new document could then be normalised with the model that most closely fits its dialect. Indeed, many words are ambiguous if we look at the entire Swiss German dialect landscape, but are disambiguated when we know which dialect is to be treated.

3.3. Part-of-speech tagging

Annotation of part-of-speech tags is a crucial step in making a corpus accessible for linguistic research through improved search facilities. In this section, we report on some experiments with automatic part-of-speech tagging, and then describe how the proposed annotation is corrected manually.

First, we manually annotate a subset of the ArchiMob corpus, consisting of 10 169 tokens in 1742 utterances, with part-of-speech tags. The annotation guidelines largely follow the ones reported by Hollenstein and Aepli (2014): they use the Stuttgart-Tübingen-Tagset (STTS) (Thielen et al., 1999) and provide some extensions to account for specific phenomena observed in Swiss German dialects. We use this set as a gold standard for testing a number of tagging models.

3.3.1. Initial experiments

In approaching the part-of-speech tagging task, we start by adapting previously developed tools and resources to our corpus.

We report here the results of the most successful tests with part-of-speech taggers trained on two different corpora. The first training corpus is TüBa-D/S, a corpus of spontaneous dialogues conducted in Standard German (360 000 tokens in 38 000 utterances). Thus, this corpus is of the same genre as our target corpus (spoken dialogue), but the two corpora belong to different linguistic varieties (standard German vs. Swiss German). The second training corpus is *NOAHs Corpus of Swiss German Dialects* (Hollenstein and Aepli, 2014), a collection of part-of-speech annotated Swiss German texts from various written sources (73 000 tokens). This corpus matches the target corpus with respect to variety (if we neglect the inter-dialectal variation), but differs in genre. All tagging experiments are carried out with BTagger (Gesmundo and Samardžić, 2012), which has shown good performance on smaller training sets.

Train	Test	% Acc.	% OOV
TüBa-D/S	Original	36.75	72.78
TüBa-D/S	Normalised	70.31	24.21
NOAH's Corpus	Original	60.56	30.72
Removed punctuation:			
TüBa-D/S	Normalised	70.68	24.21
NOAH's Corpus	Original	73.09	30.72

Table 2: Results of the part-of-speech tagging experiments. *% Acc.* reports tagging accuracy, and *% OOV* reports the percentage of words in the test set that are unknown to the tagger.

Table 2 sums up the part-of-speech tagging experiments. The tagger trained on TüBa-D/S – unsurprisingly – performs badly when applied directly to the original ArchiMob corpus: 72.78% of the words are unknown to the tagger, since they are dialect words that differ from the ones used in Standard German. This problem can be remedied by using the normalised word forms as tagger input. In this case, the proportion of unknown words goes back to 24.21%, and the accuracy is almost doubled (from 36.75% to 70.31%).

NOAH's Corpus contains Swiss German data, so that the normalised word forms are not required for tagging. The tagger trained on NOAH's Corpus yields 60.56% of accuracy with a proportion of 30.72% unknown words and thus performs a bit less well than the TüBa-D/S tagger. The higher number of unknown words is mainly due to the smaller training corpus (about one fifth of TüBa-D/S) and to the fact that parts of NOAH's Corpus are written in different dialects than the ArchiMob texts.

When inspecting the results of the two taggers, we found that the last word of each utterance had a high likelihood of being tagged as `$.` (period), because both training corpora contain syntactic punctuation. This is unwanted as the ArchiMob utterances do not contain sentence boundaries. We therefore trained new taggers with modified versions of the two training corpora where all tokens annotated with `$,` or `$.` were removed. While this removal did not have a sensible impact on the accuracy of the TüBa-D/S tagger, the accuracy of the NOAH tagger rose to 73.09%, outperforming the TüBa-D/S tagger.

The results of the tests indicate that using less training data from a non-matching genre in the same variety gives better results than using more data from a matching genre in a different variety, all the more so as certain genre-specific characteristics like punctuation can easily be adapted.

Note that the additional tags introduced in NOAH's Corpus to account for morphosyntactic particularities of Swiss German dialects help produce better results. Indeed, 2.45% of tokens in the gold standard are tagged with one of the additional tags; the NOAH tagger provides 68.05% accuracy on these tokens, whereas the TüBa-D/S tagger, having not

Round	1	2	3	4
Accuracy (%)	79.99	84.08	88.55	90.09

Table 3: The increase of the PoS tagging performance through four rounds of bootstrapping.

seen these tags in the training data, gets them all wrong.

3.3.2. Bootstrapping

The best accuracy score of 73.09% reached using the existing resources is far below an acceptable level. Our ultimate goal is to reach gold standard annotation quality for the entire ArchiMob corpus. Therefore, we apply the same bootstrapping procedure as for normalisation to correct the automatic output and increase the in-domain training set: in each round of bootstrapping, one document is automatically tagged, manually corrected, and added to the training data for the next round.

Table 3 shows the increase in the performance of the PoS tagger after the first four rounds of bootstrapping. We can see that the increase is substantial in the first three rounds, but starts to slow down in the fourth round. The performance obtained after the four steps allows for a relatively fast automatic correction. Improving the performance, however, asks for a more sophisticated approach that is yet to be developed.

4. Corpus access and formats

We provide online look-up using corpus query engines. After considering suitable engines, we decide to use three available engines: Sketch Engine (Kilgarriff et al., 2014) with the most convenient features but with the access restricted by a commercial licence, NoSketch (Rychlý, 2007) with free access but fewer search options, and ANNIS (Krause and Zeldes, 2014), an engine with free access suitable for advanced users working with complex queries.

In addition to online look-up, we provide an XML archive for download. The XML format of the documents in the archive is the base for producing the formats required by the corpus query engines. The current point of access to the corpus is its web page (`http://www.spur.uzh.ch/en/departments/korpuslab/Research/ArchiMob.html`), but we consider integrating it in a larger infrastructure (such as CLARIN).

The data are stored in three types of files:

- **Content files** contain the text of transcriptions marked with XML.
- **Media files** contain the alignment between transcribed text and the corresponding video documents.
- **Speaker files** contain 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 id="d1007-u88" who="s2" start="1007_TLI_71">
```

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

```
<w id="..." normalized="einest" POS="ADV"> 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 vocal="..."/>), 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 id="..."> zw </del>
<w id="..." normalized="zwei" POS="CARD"> zwee </w>

<w id="..." normalized="wo" POS="KOUS">won</w>
<w id="..." normalized="ich" POS="PPER">ch</w>
<pause vocal="eh" />
<w id="..." normalized="ja" POS="ITJ">ja</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. We currently do not use any mechanism for automatic inclusion of this information into the content files.

The source video documents are available on request.

5. Conclusion

In this paper, we present a general-purpose corpus of spoken Swiss German, based on manual transcriptions of video recordings. The transcriptions are aligned with the sound source at the level of segments 4-8 seconds long, which makes it suitable for training speech-to-text systems. In order to deal with writing inconsistency, inter-dialectal variation and intra-speaker variation, we add an annotation level with normalised tokens. The corpus is also annotated with part-of-speech tags. We have presented experiments to partially automate both the normalisation and part-of-speech tagging tasks, using manually annotated training data for the former and taggers trained on existing corpora for the latter. The resource presented here is freely available for research.

6. 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, Hana Ruch, Franziska Schmid, Fatima Stadler, Janine Steiner-Richter, Phillip Ströbel, Simone Ueberwasser, Alexandra Zoller.

7. Bibliographic References

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

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

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

Rudolf Hotzenköcherle, et al., editors. (1962–1997). *Sprachatlas der deutschen Schweiz*. Francke, Bern.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

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

Kolly, M.-J. and Leemann, A. (2015). Dialäkt Äpp: communicating dialectology to the public – crowdsourcing dialects from the public. In Adrian Leemann, et al., editors, *Trends in Phonetics and Phonology. Studies from German-speaking Europe*, pages 271–285. Peter Lang, Bern.

Krause, T. and Zeldes, A. (2014). Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*.

Rychlý, P. (2007). Manatee/Bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno: Masaryk University.

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

Scherrer, Y. and Rambow, O. (2010). Natural language processing for the Swiss German dialect area. In *Proceedings of KONVENS 2010*, Saarbrücken.

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

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

Friedrich Staub, et al., editors. (1881–). *Schweizerisches Idiotikon : Wörterbuch der schweizerdeutschen Sprache*. Huber, Frauenfeld.

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