Future Actions for Swiss German — Workshop Results at SwissText 2018

Tanja Samardi
Language and Space Lab
University of Zurich
tanja.samardzic@uzh.ch

Mark Cieliebak
School of Engineering
Zurich University
of Applied Sciences
mark.cieliebak@zhaw.ch

Jan Milan Deriu
School of Engineering
Zurich University
of Applied Sciences
jan.deriu@zhaw.ch

Abstract

The goal of this workshop was to initiate collaborations among companies and academic institutions for developing Swiss German resources and activities. The need for such an initiative is created by a growing interest for applying automatic text processing technologies to Swiss German, which takes place in the context of particularly scarce data sets. We have considered potential modes for a collaborative data development and management. The outcome of the workshop are defined common interests, priorities, and the first steps in future synchronised efforts.

1 Introduction

Automatic processing of Swiss German has long been regard as not needed, as standard German is regularly used in public communication in Switzerland. This view, however, has recently changed following the increased presence of local varieties in public communication (mostly on the Internet). This brought several companies and academic institutions to start working on automatic processing of Swiss German. This has resulted in the development of initial data sets that can be used for training models for automatic speech and text processing. These data sets are, however, scattered across different institutions that produced them and not easily accessible to the researchers outside the host institutions. On the other hand, each data set individually is too small to allow good performance on tasks such as speech recognition, translation, or normalisation. All involved parties need more data and would greatly benefit from an exchange and future joint development. The goal of this workshop was to bring together researchers working on Swiss German in companies and in academic institutions in order to identify the common needs and modes of future collaborations.

In: Mark Cieliebak, Don Tuggener and Fernando Benites (eds.): Proceedings of the 3rd Swiss Text Analytics Conference (SwissText 2018), Winterthur, Switzerland, June 2018

2 Overview of the Current Activities

As part of the preparations for the workshop, we have conducted an informal survey among the researchers and institutions we knew were interested or already working on Swiss German automatic processing. We asked the contacted persons to share with us an overview of their data sets, tools, and general activities related to Swiss German. We have received ten responses, coming from the following organisations:

- Institute of Computational Linguistics, University of Zurich
- Swiss Re
- School of Business and Engineering Vaud
- Swisscom
- Institute of Applied Information Technology, Zurich University of Applied Sciences
- School of Applied Linguistics, Zurich University of Applied Sciences
- Spitch AG
- University of Helsinki
- Slowsoft

Based on these responses and our own insight, we have composed an initial overview of the existing

CEUR-WS.org/Vol-2226/paper12.pdf

Institution	Group/Individual
School of Business and Engineering Vaud (HEIG-VD)	Andrei Popescu-Belis
The Idiap Research Institute (Idiap)	Walliserdeutsch
Schweizerisches Idiotikon	
University of Geneva (UniGe)	Manuela Schnenberger, Eric Haeberli
University of Helsinki	Yves Scherrer
University of Zurich (UZH)	Language and Space Lab, German Department, Romance Department, Institute of Computational Linguistics
Zurich University of Applied Sciences (ZHAW)	Text Analytics and Dialogue Systems Group

Table 1: Research groups in academic institutions in alphabetic order

Company
Recapp
Spitch
SpinningBytes
Slowsoft
Swisscom

Table 2: Companies working on/with Swiss German

data, tools, and addressed processing tasks. This information is presented in the remainder of this section. While we did our best to collect as much information as possible in the present moment, this review is not to be regarded as an exhaustive inventory, but rather as a first step towards a complete inventory that will be developed through collaborative work.

2.1 The List of Institutions

In the research on automatic processing of Swiss German both practical and scholarly sides are equally pronounced. In order to get the standard natural language processing work for Swiss German, we need to understand and address the details of its particular and complex usage practices. There is thus a considerable overlap between the work on developing end-user applications, primarily done in the companies, design of algorithmic solutions, typically at applied universities, and data-driven study of linguistic variation in Swiss German, primarily performed in academic institutions.

We have identified several groups inside academic institutions where some work relevant to the automatic processing of Swiss German is taking place. They are listed in Table 1. For Swiss institution, we give in the parentheses the corresponding abbreviations, which are easily recognisable to the Swiss audience.

An additional institution where some work on Swiss German is likely to take place is ETH Zurich, but we have not established a relevant contact up to this point.

Considerable efforts have already been invested in processing Swiss German in the companies listed in Table 2. Additionally, one company, Telepathy Labs, is associated with a published piece of work on Swiss German, but their engagement is yet to be confirmed.

2.2 Data Sets

Here we list the Swiss German data sets produced and made available by the groups listed above and some other groups who worked on Swiss German in the past. This list relies on three publications that appeared in the proceedings of the Language Resources and Evaluation Conference (LREC 2018), all providing informative overviews of the state of the art:

- SB-CH: A Swiss German Corpus with Sentiment Annotations
  R. Grubenmann, D. Tuggener, P. Von Dniken, J. Deriu, M. Cieliebak
- Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German
  P.-E. Honnet, A. Popescu-Belis, C. Musat, M. Baeriswyl
- Strategies and Challenges for Crowdsourcing Regional Dialect Perception Data for Swiss German and Swiss French
  J.-P. Goldman, S. Clematide, M. Avanzi, R. Tandler

We divide the data sets into two major types: text corpora (Table 3) and lexica (Table 4). For each listed item, we specify the institution or the group that developed it and the most important characteristics of the data set. The column "Text" specifies whether the resource contains text. "Sound" whether it contains recorded speech (both are present when sound recordings are transcribed). "Norm/Trans" specifies whether there is word level normalisation of writing or full translation to standard German. We join these two features together because normalisation typically involves standard German writing applied to Swiss German. The difference between the two is that normalised text is not necessarily proper standard German in the sense of orthography, grammar and style. The last column "PoS" specifies whether the text is annotated with part-of-speech tags. We list institution name where it is clear which institution is responsible for the resources, otherwise, we provide the information about its authors.

There are two specific remarks regarding Table 3. First, note that two corpora contain additional annotation: the corpus SB-CH sentiment and the UniGe corpus syntax. Second, the data used in the project "din dialkt" are often taken from already existing resources, which means that there are considerable overlaps between this set and other known sources.

Regarding Table 4, we use parentheses "( )" to signal two remarks. First, mapping to standard German referred to in the column "Norm/Trans" exists in the resources built by the Idiotikon team, but it is not encoded in the same way as in the corpora. This applies to the part-of-speech information too: while there is information on the word types in the dictionary, these codes do not follow usual German tag sets. Second, the data used for the projects by Leemann et al. are drawn from other sources (SDS), resulting in a considerable overlap. Also, mapping to standard German and part-of-speech information is likely inherited from SDS.

Overall, Table 3 and 4 show that most available data sets come from academic institutions. While companies can be expected to have developed their own resources too, descriptions of these resources are yet to be shared.

2.3 Processing Tasks and Tools

Natural language processing can potentially involve many different tasks for which specific tools are developed. We list here those tasks that have been addressed for Swiss German or that are mentioned as current activities in the contacted institutions. Assuming the view of natural language processing as a pipeline, or a stream, we divide the tasks into two groups: upstream tasks (Table 5) and end-user tasks (including annotation tools, Table 6). The output of the first group of tasks is not necessarily visible to the end user, but rather used as input to the end-user tasks.

The parentheses in these two tables indicate different remarks. In Table 5, they are used to specify the institution where the work on the given task is performed. In Table 6, the parentheses indicate that it is not clear at this point whether the tasks are attempted specifically for Swiss German, since most of the institutions develop their applications for multiple languages. Although it is not surprising, it is interesting to note that Table 5 lists mostly academic institutions, while companies are more involved in tasks listed in Table 6.

3 Future Actions

Most of the groups listed in our overview were represented at the workshop, which was generally very well attended. The live discussion that followed the overview of the current state of the resources and tools addressed the following points.

Comments on the overview

The comments from the audience on the presented review showed that several participants were involved in projects on speech recognition with Recapp, not covered by the overview. They also pointed out the resources developed by Slowsoft (transcribed Swiss German sentences, pronounced by one speaker), that were not listed in the overview due to a miscommunication. Other potential sources of data were mentioned, such as SRF subtitles that are in standard German, but aligned with Swiss German sound source.

Identifying common needs and priorities for further development

The discussions on the common needs revolved around the question of writing for Swiss German

Project	Who	Text	Sound	Norm/Trans	PoS
ArchiMob	UZH+Spitch	✓	✓	✓	✓
BE-Novel	Honnet et al.	✓		✓	
NOAH	UZH	✓			✓
Phonogram	UZH	✓	✓	✓	
SB-CH (Sentiment)	ZHAW+SpinningBytes	✓			
sms4science	UZH+Swisscom	✓		✓	✓
walliserdeutsch	Idiap	✓	✓	✓	
Wil corpus (Syntax)	UniGe	✓	✓		✓
WUS (WhatsApp)	UZH	✓		✓	✓
din dialkt	UZH	(✓)	(✓)	(✓)	(✓)
In progress	UZH	✓	✓		

Table 3: An initial inventory of Swiss German text corpora available for training processing tools.

Project	Who	Text	Sound	Norm/Trans	PoS
Swiss German Atlas (SDS)	Schweizerisches Idiotikon	✓		(✓)	(✓)
Swiss German Dictionary	Schweizerisches Idiotikon	✓		(✓)	(✓)
Dialkt pp	Leemann et al.[1]	(✓)		(✓)	(✓)
Voice pp	Leemann et al.[2]	(✓)	(✓)	(✓)	(✓)
Pronunciation	Spitch	✓	✓		
BE-Lexicon	Honnet et al.	✓		✓	
ZH-Lexicon	Honnet et al.	✓		✓	

Table 4: An initial inventory of Swiss German lexical resources (potentially) relevant to automatic processing.

Task	Who
Speech recognition	Spitch, UZH
Normalisation	ArchiMob, SMS, WhatsApp (UZH)
Anonymisation	Swisscom, UZH
Morphology (finite-state)	Scherrer PhD thesis (UniGe), Baumgartner MA thesis (UZH)
PoS	ArchiMob, NOAH, SMS, WhatsApp (UZH)
Syntax	Forst MA thesis (UniL), Aepli MA thesis (UZH)

Table 5: Upstream tasks that have been attempted for Swiss German

Task	Who
Active learning	Swisscom, UZH
Dialect identification	Swisscom, UZH, ZHAW
Normalisation	UZH
Sentiment annotation	Swisscom, SpinningBytes
Sentiment classification	(Swisscom), SpinningBytes, ZHAW, (Spitch)
Speaker identification	(Swisscom), Spitch
Speech Synthesis	(Slowsoft), (Swisscom), (Spitch)
Transcription	Swisscom-AILA

Table 6: End-user applications and annotation tools that are being developed for Swiss German

texts. As there is no official standard, the text is likely to be written either in a non-standard way (as in the user-generated content) or in standard German (as in the case of SRF transcriptions, for instance). Mapping speech to a standard writing came out as a common need. Defining and implementing a common writing standard seems to be one of the potential topics for collaboration. The discussion also showed that, to define other common needs, we would need to analyse real use scenarios and identify the tasks based on them.

Modes of collaboration

The discussion on this topic showed that there is an interest for collaboration, but that it will take considerable work in order to make it work. One obstacle is the fact that companies typically work with sensitive data that cannot be shared. Another problem is that sharing data requires additional work (such as anonymisation, detailed documentation). With a clear idea of the potential benefits, the groups working on Swiss German might be ready to invest more effort in order to facilitate collaboration.

Funding possibilities

For the moment, the foundation InnoSuisse seems like the best choice for submitting proposals. It enables developing solutions for an identified need without a concrete business plan. Other foundations targeting the exchange between academia and companies can be considered.

Next steps

Based on the previous discussions, we have defined the following actions as the next steps towards establishing a collaborative network for Swiss German:

- Start collaborating through proposing the first InnoSuisse grant as soon as possible.
- Organise shortly a follow-up workshop where the ideas for the InnoSuisse proposal will be sketched (2/3 participants were interested in attending such a workshop).
- Elaborate and share a detailed inventory of the existing resources and tools, including the information on the conditions of use and data samples.
- Formulate processing tasks based on use cases.

This plan can be considered the main outcome of the workshop, together with the established contacts and identified commitments to work together on synchronising efforts invested in processing Swiss German.