Compiling a Large Swiss German Dialect Corpus

Manuela Weibel
Schweizerisches Idiotikon
Auf der Mauer 5
8001 Zürich
manuela.weibel@idiotikon.ch

Muriel Peter
Schweizerisches Idiotikon
Auf der Mauer 5
8001 Zürich
muriel.peter@idiotikon.ch

Abstract

The *Swiss German Dialect Corpus (Schweizer Mundartkorpus CHMK)* is an initiative launched by the Swiss German dictionary *Schweizerisches Idiotikon*. It is an unbalanced, opportunistic corpus and the largest dialect corpus for Swiss German to date. The corpus will be accessible through a query engine and, in part, as an open-source XML corpus. In this paper we provide an overview of the concept, workflow, and challenges of compiling a corpus for a non-standard linguistic variety.

1 Introduction

1.1 The language situation in German-speaking Switzerland

Switzerland has four official languages (German, French, Italian, and Romansh) with German being the most widely spoken (more than 60 percent[1]). Within German-speaking Switzerland, two varieties of the same language co-exist and are used in separate and distinct situations: Swiss German and standard German. Such a situation is widely referred to as a diglossia (cf. Ferguson, 1959; Rash, 1998; Christen, 2019; Christen and Schmidlin, 2019). One variety, in this case Swiss German, is commonly used for everyday, mostly spoken communication. It is usually not codified and is associated with informal situations. The second variety, standard German, is highly codified and used in formal settings such as school, political debates, information programmes on national radio or television as well as for written texts (Christen and Schmidlin, 2019, p. 208).

The Swiss German dialects are part of the Alemannic dialect group and form a continuum: there are no clear-cut boundaries between the different dialects; some phenomena occur in more than one dialect, while others are unique. Moreover, due to the small-scale nature of the linguistic areas, a lot of variation occurs within Swiss German dialects. And even though Swiss German is increasingly being used for written communication, there is no standardised orthography. Attempts to introduce a standard for writing Swiss German dialects have had varying degrees of success (cf. Rash, 1998; Siebenhaar, 2013; Christen and Schmidlin, 2019). One such attempt is the "Schwyzertütschi Dialäktschrift", introduced by Eugen Dieth in 1938 and updated by Christian Schmid-Cadalbert in 1986. It applies to all Swiss German dialects and is widely used by linguists and dialectologists today (cf. Scherrer et al., 2019). However, as none of the standards have ever been taught at school, they have not been implemented by a significant number of people. There is no political intent for establishing a standard Swiss German, nor is there a need for it, as Swiss German speakers may make use of standard German in order to be understood in other German-speaking countries. These three factors – the dialect continuum, the small-scale linguistic landscape, and the lack of a standardised orthography – result in a large degree of variation in written Swiss German (cf. Christen, 2019).

German-speaking Switzerland has a rich tradition of dialect literature. Beyond that, its use in written communication was limited in the past. However, with the development of text message

[1] Federal Statistical Office: "Main languages of the permanent resident population, 1970-2017", published: 21.02.2019, https://www.bfs.admin.ch/bfs/de/home/statistiken/kataloge-datenbanken/grafiken.assetdetail.7466560.html.

services and social media, the number of texts written in Swiss German has increased significantly over the last 20 years (Samardžić et al., 2015; Christen and Schmidlin, 2019).

1.2 Intentions of the project

Possibilities concerning natural language processing (NLP) for Swiss German have long been restricted due to an insufficient number of available texts. With the rise of Swiss German text resources, Switzerland's language technology research has recently shifted its focus: first efforts towards building dialect corpora and developing NLP tools for Swiss German have been made since 2009 (see subsection 2.1 for a more detailed view on related research). With the *Swiss German Dialect Corpus (Schweizer Mundartkorpus CHMK)*, we intend to further facilitate this development by building the largest integral corpus of Swiss German dialects so far. The *Schweizerisches Idiotikon* is able to provide the required human and material resources for the compilation of such a corpus – most importantly, an extensive library of dialect literature.

As a research institution that is dedicated to the documentation of Swiss German dialects, the *Schweizerisches Idiotikon* intends to provide a platform for existing and future Swiss German dialect corpora.

2 Related Work

2.1 Language technology for Swiss German

The *Swiss SMS corpus* compiled by Dürscheid and Stark (2011) features nearly 26,000 short messages in different languages and dialects with more than 40 percent of the messages written in Swiss German. The corpus is part-of-speech tagged. Normalisation[2] was conducted by means of interlinear glossing with a specifically developed annotation tool (Ruef and Ueberwasser, 2013) and following a continuously updated set of annotation guidelines (cf. Ueberwasser, 2013).

NOAH's corpus by Hollenstein and Aepli (2014) is a comparably small collection of manually part-of-speech tagged Swiss German texts. Hollenstein and Aepli adapted the Stuttgart-Tübingen-Tagset (STTS), a part-of-speech tagset widely applied in NLP for standard German, accounting for the morphosyntactic particularities of Swiss German dialects.

The *ArchiMob corpus* by Samardžić et al. (2016) provides transcriptions of spoken Swiss German. Samardžić et al. (2015) conducted a range of experiments in order to automate annotation steps: they examined different methods based on machine translation in order to normalise the corpus. Furthermore, part-of-speech tagging was applied semi-automatically, based on the tagset established by Hollenstein and Aepli (2014).

For the project *What's up, Switzerland?*, Ueberwasser and Stark (2017) collected more than one million WhatsApp messages. Nearly half of the messages are written in Swiss German. The corpus is due for publication in spring 2020.

It is our goal to integrate existing corpora into the *CHMK* wherever licences are compatible and we plan on using the part-of-speech tagset provided by Hollenstein and Aepli (2014).

2.2 Schweizer Textkorpus CHTK

The *Swiss Text Corpus CHTK* is a balanced reference corpus of the written standard language in German-speaking Switzerland in the 20th and 21st centuries. It was established in 2000 at the University of Basel, with the aim of assembling a corpus of standard German texts from Switzerland and ensuring its accessibility and continuation. In 2014, the corpus was transferred to the *Schweizerisches Idiotikon*, where it has been maintained ever since. Various insights gained from this corpus can be applied to the *CHMK*. A short overview of the *CHTK* will contrast its selection criteria with those we have defined for the new dialect corpus (see subsection 3.2).

The texts in the *CHTK* were chosen based on different criteria regarding form, content, and time of publication. Hard criteria, which had to be met, and soft criteria were established. Apart from the language in which they were written (standard German) and the time of publication, further hard criteria included form and content of the texts. Four categories of work were defined, which had to be represented evenly: fiction, non-fiction, functional texts, and journalistic texts.

Due to the limited number of texts available for the 20th century, soft criteria such as the author's regional origin within Switzerland and their gender could not always be taken into consideration. For further information on the *CHTK* see Bickel et al. (2009).

[2] Word normalisation describes the process of identifying multiple forms of a word and assigning a single normal form.

genre	books	~ number of pages (in Swiss German)	average token count per page	~ number of tokens (in Swiss German)
prose	405	66,800 (62,700)	350	23,380,000 (21,945,000)
poetry	129	15,000 (13,600)	60	900,000 (816,000)
drama	40	2,400 (2,000)	200	480,000 (400,000)
mixed	48	7,700 (7,200)	300	2,310,000 (2,160,000)
total	622	91,700 (85,500)		27,070,000 (25,321,000)

Table 1: Expected size of *CHMK* (first release).

3 Swiss German Dialect Corpus CHMK

3.1 Application

The *Swiss German Dialect Corpus* will allow a set of practical applications. A corpus query engine will provide an interface for linguistic research. The engine will build upon the knowledge and expertise gained when working on the *CHTK*[3]. In addition, the corpus will serve as a base for the lexicographic work on the Swiss German dictionary *Schweizerisches Idiotikon*. Finally, all copyright-free texts of the corpus will be made fully accessible in XML format and under an open-source licence. This way, we intend to provide a tool for the enhancement of language technology research for Swiss German dialects.

3.2 Selection criteria

In contrast to the previously compiled reference corpus *CHTK*, we have decided that the new dialect corpus will be an unbalanced corpus of Swiss German texts. This decision is mainly based on the fact that an equal distribution of text genres and dialects is difficult to ensure for a low-resource language such as Swiss German, and would pose an unnecessary constraint on the number of eligible texts per criterion.

There has been a recent increase of available texts written in Swiss German, one important source for non-fictional Swiss German texts being the Alemannic Wikipedia[4]. Nevertheless, the list of represented text genres shows a considerable bias towards fiction and poetry: technical texts such as medical essays or instruction manuals written in Swiss German are difficult to obtain.

Due to a lack of availability, we also have to abstain from weight criteria concerning authors' gender and dialect distribution. As a result, the *CHMK* selection criteria forgo any weighting and are reduced to the following:

1. The texts need to be written in Swiss German.
2. The texts need to be from 1800 AD or later.
3. The dialect must be clearly indentifiable, i.e. the canton or region must be known.

While the limited number of criteria ensures a larger number of eligible texts, it may simultaneously lead to an overrepresentation of certain dialects and an unbalanced gender distribution. It is therefore important that users of our corpus take into account that we do not intend to represent the linguistic reality of German-speaking Switzerland but to gather and provide as much data as possible.

3.3 Processing

The *Swiss German Dialect Corpus* will potentially serve many different purposes. Thus, it is crucial that we meticulously document the metadata of the gathered sources; author biographies, text categorisation and dialect information will later enhance a wide range of linguistic analyses. Detailed metadata also allow for more specific machine learning training, i.e. when it comes to training a language model for a certain dialect. Last but not least, well-documented metadata will simplify the creation of copyright-dependent and other subcorpora.

So far, we have collected and scanned over 600 books, adding up to over 90,000 pages. Since optical character recognition (OCR) has not yet been performed, we can only estimate the actual number of tokens. Taking into consideration different average word counts depending on the text genre, we expect our scanned texts to contain a total of over 27 million tokens (see Table 1).

[3]The *CHTK* corpus query tool is based on the open source search engine ddc-concordance (http://www.ddc-concordance.org/).

[4]https://als.wikipedia.org.

When estimating the potential number of Swiss German tokens, we deduct 10 pages per book, accounting for illustrations, titles and index pages, as well as other sections generally written in standard German (introduction, epilogue, author biography, bibliography). This results in a total of approximately 25.3 million Swiss German tokens.

In order to acquire more recent dialect literature, we have downloaded 900 individual webpages, counting approximately 500,000 tokens in total.

Before any further texts are gathered, the data obtained so far will be processed and linguistically annotated. This will allow for an earlier publication of a first release. Moreover, repeating the workflow for further releases will lead to an additional refinement of the process.

4 Prospect

In a next step, the scanned texts will be converted into digital text using OCR. For this purpose, we have teamed up with a group of developers in Würzburg (Germany), using their system *OCR4all*[5]. *OCR4all* is an open-source software which combines state-of-the-art OCR components and continuous model training into a full workflow. Primarily developed to analyse historical printings, the system not only performs well on modern fonts, it also outperforms commercial state-of-the-art tools when applied to 19th century Fraktur scripts (cf. Reul et al., 2019). Given that almost a sixth of our scanned books are, in fact, printed in Fraktur typeface, this is an important asset for our project. An included module allows for the correction of errors and training of new recognition models. Tests will be carried out to evaluate the cross-dialectal range of a trained model and assess the need for a dialect classifier.

Given the relatively large size of our corpus, it is important that we automate as many of the processing steps as possible. For this reason, we intend to apply automated methods for part-of-speech tagging and normalisation as assessed by Samardžić et al. (2015).

The absence of a writing standard for Swiss German paired with large lexical and phonological differences across German-speaking Switzerland result in substantial orthographic inconsistencies. These can be observed, not only on an inter-dialectal level or between different writers, but even on an intra-writer level. When providing a corpus query engine for non-standard linguistic varieties such as Swiss German, it is therefore crucial that the data is normalised. Both the *ArchiMob Corpus* and the *Swiss SMS Corpus* have developed their own normalisation guidelines. In an effort to harmonise existing and future corpora, we plan to establish a normalisation standard for Swiss German and apply it to our corpus. In case of compatible licences, we will also apply it to existing corpora and integrate these into our first release, thus converting the *Swiss German Dialect Corpus* to a central platform for all dialect corpora for Swiss German.

[5]https://www.uni-wuerzburg.de/zpd/ocr4all.

References

Hans Bickel, Markus Gasser, Annelies Häcki Buhofer, Lorenz Hofer, and Christoph Schön. 2009. Schweizer Text Korpus – Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten. *Linguistik Online*, 39(3):5–31.

Helen Christen. 2019. Alemannisch in der Schweiz. In Hanna Fischer and Brigitte Ganswindt, editors, *Deutsch: Sprache und Raum-Ein internationales Handbuch der Sprachvariation*, volume 30 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 246–279. Walter de Gruyter GmbH & Co KG.

Helen Christen and Regula Schmidlin. 2019. Die Schweiz. Dialektvielfalt in mehrsprachigem Umfeld. In Rahel Beyer and Albrecht Plewnia, editors, *Handbuch des Deutschen in West- und Mitteleuropa: Sprachminderheiten und Mehrsprachigkeitskonstellationen*, pages 193–244. Narr Francke Attempto Verlag.

Eugen Dieth. 1938. *Schwyzertütschi Dialäktschrift: Leitfaden nach den Beschlüssen der Schriftkommission der Neuen helvetischen Gesellschaft, Gruppe Zürich*. O. Füssli.

Eugen Dieth and Christian Schmid-Cadalbert. 1986. Schwyzertütschi Dialäktschrift. *Sauerländer, Aarau*, 2.

Christa Dürscheid and Elisabeth Stark. 2011. *sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland*, pages 299–320.

Charles A. Ferguson. 1959. Diglossia. *WORD*, 15(2):325–340.

Nora Hollenstein and Noëmi Aepli. 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*, pages 85–94,

Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Felicity J. Rash. 1998. *The German language in Switzerland: multilingualism, diglossia and variation*, volume 3. Peter Lang Pub Inc.

Christian Reul, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner, and Frank Puppe. 2019. OCR4all—An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. *Applied Sciences*, 9:4853.

Beni Ruef and Simone Ueberwasser. 2013. The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages. In Marcos Zampieri and Sascha Diwersy, editors, *Non-standard Data Sources in Corpus-based Research*, ZSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln, pages 61–68. Shaker Verlag, Aachen.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. Archimob - a corpus of spoken swiss german. In *Language Resources and Evaluation (LREC 2016)*, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 4061–4066. s.n.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2015. Normalising orthographic and dialectal variants for the automatic processing of Swiss German. Proceedings of the 7th Language and Technology Conference.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769. Publisher: Springer.

Beat Siebenhaar. 2013. Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats.

Simone Ueberwasser. 2013. Non-standard data in Swiss text messages with a special focus on dialectal forms. In Marcos Zampieri and Sascha Diwersy, editors, *Non-standard Data Sources in Corpus-based Research*, ZSM-Studien, Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit der Universität zu Köln, pages 7–24. Shaker Verlag, Aachen.

Simone Ueberwasser and Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5):105–126.