ArchiMob: A multidialectal corpus of Swiss German oral history interviews

Yves Scherrer, University of Helsinki
Tanja Samardžić, University of Zurich

Although dialect usage is prevalent in the German-speaking part of Switzerland, digital resources for dialectological and computational linguistic research are difficult to obtain. In this paper, we present a freely available corpus of spontaneous speech in various Swiss German dialects. It consists in transcriptions of video interviews with contemporary witnesses of the Second World War period in Switzerland. These recordings were produced by an association of Swiss historians called Archimob[1] about 20 years ago. More than 500 informants originating from all linguistic regions of Switzerland (German, French and Italian) and representing both genders, different social backgrounds, and different political views, were interviewed. Each interview is 1 to 2 hours long. In collaboration with the University of Zurich, we have selected, processed and analyzed a subset of 43 interviews in different Swiss German dialects.

The goal of this contribution is twofold. First, we describe how the documents were transcribed, segmented and aligned with the audio source and how we make the data available on specifically adapted corpus query engines. We also provide an additional layer in which each transcribed word form is associated with a normalized form. This normalization reduces the different types of variation (dialectal, speaker-specific and transcriber-specific) present in the transcriptions; the normalization language resembles standard German. We formalize normalization as a machine translation task, obtaining up to 90% of accuracy (Scherrer & Ljubešić 2016).

Second, we show through some examples how the ArchiMob resource can shed new lights on research questions from digital humanities in general and dialectology and history in particular:

- Thanks to the normalization layer, dialect differences can be identified and compared with existing dialectological knowledge.
- Using language modelling, another technique borrowed from language technology, we can compute distances between texts. These distance measures allow us to identify the dialect of unknown utterances (Zampieri et al. 2017), localize transcriber effects and obtain a generic picture of the Swiss German dialect landscape.
- Departing from the purely formal analysis of the transcriptions for dialectological purposes, we can apply methods such as collocation analysis to investigate the content of the interviews. By identifying the key concepts and events referred to in the interviews, we can assess how the different informants perceive and describe the same time period.

References

Tanja Samardžić, Yves Scherrer & Elvira Glaser (2016): *ArchiMob - A corpus of Spoken Swiss German*. Proceedings of LREC 2016, 4061-4066, Portorož, Slovenia.

Yves Scherrer & Nikola Ljubešić (2016): *Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation*. Proceedings of KONVENS 2016, 248-255, Bochum, Germany.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, Noëmi Aepli (2017): *Findings of the VarDial Evaluation Campaign 2017*. Proceedings of the VarDial 2017 Workshop, EACL, Valencia, Spain.

[1] Archimob stands for *Archives de la mobilisation*, i.e. archives of the mobilization period.