The Swiss SMS Corpus

The Project

The Swiss SMS corpus is one of the results of a project funded by the Swiss National Science Foundation between 2011 and 2014 and directed by Prof. Elisabeth Stark. Other results of the project are six dissertations as well as an abundance of student papers and publications. An overview over the whole project with can also be found in the SNSF research database P3.

The corpus

The Swiss SMS corpus consists of 25'947 SMS (~650'000 tokens), which were sent in by the Swiss public in 2009/2010. Of all SMS, 41% are in Swiss German (dialect), 28% in non-dialectal German, 18% in French, 6% in Italian, and 4% in Romansh. More information about the corpus can be found in the section facts and figures.

Using the corpus

These data are freely available for bonafide academic research (CC-NY-NC), but not for commercial use. If you use the corpus, you agree to our conditions, i.e. to:

- Not use the data for commercial use, i.e. only for bonafide research
- Quote the source of the data as "Swiss SMS corpus" with the source as shown in the footer of this document and with a link to https://sms.linguistik.uzh.ch

If you need help browsing the corpus, please check the chapter Browsing.

Since the corpus is available on the same platform as the data from the sister-project What's up, Switzerland?, please keep in mind that only the following sub-corpora contain SMS data, while the other sub-corpora are built up of WhatsApp messages.

- deu-rftagged: non-dialectal German data tagged with RF-Tagger
- deu-tagged: non-dialectal German data tagged with TreeTagger
- fra-tagged: French data tagged with TreeTagger
- gsw-rftagged: Swiss German data where the normalized data was tagged with RF-Tagger
- gsw-tagged: Swiss German data where the normalized data was tagged with TreeTagger
- ita-tagged: Italian data taggend with TreeTagger
- roh: Romansh data

For more information about the WhatsApp corpus, please consult the according documentation.

How to quote

Quoting the corpus

Stark, Elisabeth; Ueberwasser, Simone; Ruef, Beni (2009-2015). Swiss SMS Corpus. University of Zurich. www.sms4science.ch

Quoting the corpus documentation

Ueberwasser, Simone (2015/2022): The Swiss SMS Corpus. Documentation, facts and figures. www.sms4science.ch

More resources that document the creation of the corpus:

Ruef, Beni/Ueberwasser, Simone (2013): The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages . In: Non-standard Data Sources in Corpus-based Research (ZSM-Studien 5). Aachen: Shaker, 61-68.

Publications that are based on the corpus

We have the full publication list on the SNSF research database P3.

Acknowledgement

This corpus would not be here without the following people/institutions, to whom we express our gratitude:

- The Swiss National Science Foundation financed the project and the dissertations over four years.
- The UZH Zurich Center for Linguistics supported us throughout the project with their IT knowledge and hosts the corpus.
- The UZH Linguistic Research Infrastructure Project hosts this project.
- Anke Lüdeling and her team allow us to use ANNIS as a browsing tool.