1. THE CORPUS

The corpus consists of 617 chats that were sent in by the Swiss population in 2014 through a fixed procedure that was communicated in the press in order to get people interested. The individual chats were checked for their permission to use them and chats that did not have it were removed. Furthermore, available demographic data were linked to the chats.

Next processing steps comprised anonymization, the annotation of a main language per chat and thus the creation of subcorpora, application of further annotations (for languages, i.e. each message was annotated for its most likely language as opposed to the chat annotation performed in the first step), part of speech annotations, normalization for part of the dialectal Swiss German data.

Our authentic WhatsApp chats were gathered in summer 2014. Not all made it into the corpus (e.g. doublets, chats or message without permission etc.). In its present form, the corpus comprises:

- Number of chats: 617
- Number of messages (with permission to be used): 763’644
- Number of tokens: 5'155'476 (without redactedQ.* (cf. Messages without permission))
- Number of emojis: 382'116

The corpus is built up of chats in all four national languages of Switzerland, i.e. Swiss German dialect, non-dialectal German, French, Italian and varieties of Romansh. In more detail, the following languages and varieties can be found in the corpus:

Available languages:

- fra: French
- ita: Italian
- roh: Any variety of Romansh
- gsw: dialectal German as used in Switzerland
- deu: non-dialectal German
- eng: English
- spa: Spanish
- sla: Any Slavic language

Romansh varieties:

- roh-ja: Jauer Romansh
- roh-sr: romontsch sursilvan
- roh-st: rumàntsch sutsilvan
- roh-sm: rumantsch surmiran
- roh-pt: rumauntsch puter
- roh-vl: rumantsch vallader
- roh-gr: rumantsch grischun