Facts and Figures: the Corpus

The corpus consists of roughly 500'000 tokens. However, counting tokens in a corpus with that many emoticons and other special characters as well as with a spelling that deviates greatly from the norm is nearly impossible. There is e.g. one participants who does not use any spaces in his SMS. His SMS consequently get counted as one single token. Thus, the figure has to be seen as an approximation.

Number of characters

The same goes for the average number of characters in the SMS: This figure can be set around 110 characters. However, some participants sent in whole protocols consisting of up to 10 SMS. Of course this behaviour raised the average.

Tokens per language

The problem becomes even bigger when trying to count the total number of tokens per language/variety. All the problems mentioned above apply here, too. Additionally, we do not really know the language of each word, but only the language of an individual SMS. Let us look at the following SMS as an example: *Sounds good;-) freu mich!!*_. This SMS is marked as both German and English, because it contains the same number of tokens in either language, i.e. because two tokens are English and two are German. In the figures below which mention tokens in German, the two words *Sounds* and good are consequently counted as German, too. But this type of problem does not only present itself in SMS that were considered as bilingual, but also in SMS with lots of nonce borrowings:

- Olla fratello!!! Come stai? Wie geht's dir so? Immer noch so lange am arbeiten wie früher? Ich hab endlich mein eigenes Restaurant und mucho travajo... 😉 aber macht mir extrem spass... 😉 allora amore, buona giornata und luegsch uf di, gäll...;-)peace

This SMS is counted as Standard German, because this language contributes the most words. However, if we quote the number of German Standard words in the corpus, we include all the Italian, Spanish, Swiss German and English words in this SMS into the total of German words in the corpus, too.

After all these warnings, which are especially important when applying statistics, let us give you some figures about the tokens per language:

- Swiss German: 275'000 tokens
- Standard German: 174'000 tokens
- Standard French: 121'000 tokens
- Standard Italian: 39'500
- Romansh: 28'000
- Italian Dialect: 1'000

We do not provide any figures for other languages/varieties, such as for Romansh varieties, because they are too small to be considered for statistics, anyway.