# SMS in the corpus

There were two collections of SMS. The first produced 23'988 SMS and took place between end October 2009 and February 2010. The second collection was only advertised on the Italian and Romansh part of Switzerland in order to produce more SMS in those two languages. It produces another 1'959 SMS, mainly in the intended languages but also in others. This collection took place between end April 2011 and July 2011.

The total number of SMS to be found in our corpus thus comes to 25'947.

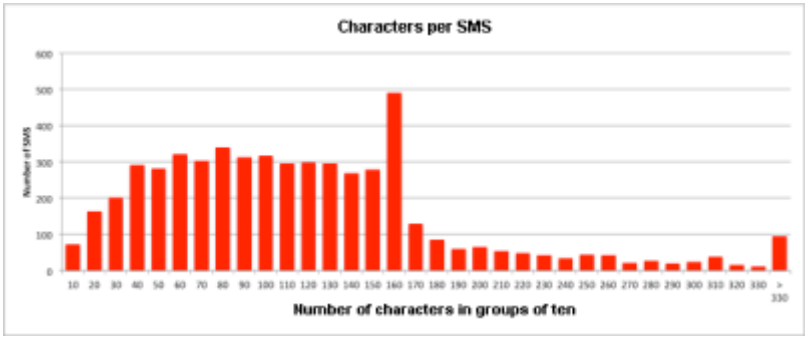
## Some statistics

The average participant sent us approx. 15 SMS. The most SMS we received from a 39-year-old man, he sent us 358 SMS. A total of 15 participants sent us more than 100 SMS each.

## Length of the sms

Some figures about the length of the SMS:

- Shortest SMS: 1 character
- Longest SMS: 2'374 characters
- Average length of the SMS (=mean): 115 characters
- Standard deviation: 84.61
- Median: 104 characters

Distribution: 


## Tokens per SMS

The following average number of tokens can be found in our SMS:

- Romansh: 68 tokens / SMS
- Non-dialectal Italian: 26 tokens / SMS
- Italian dialect: 22 tokens / SMS
- Non-dialectal German: 24 tokens / SMS
- German dialect: 26 tokens / SMS
- Non-dialectal French: 26 tokens / SMS
- French dialect: 33 tokens / SMS

Please consider our warnings about counting tokens in SMS before working with these figures.