The collection

Data collection

The data were collected in two stages. A first collection took place Nov 2009 - Feb 2010. Because this collection did not produce enough SMS for research in Italian and Romansh, a second collection took place between May and July 2011 and produced SMS mainly in those two languages. The two collections are now fully integrated and appear as one single corpus. In your everyday work with the corpus, you will not know whether a specific SMS was collected in the first or in the second round. A small difference can still be seen in the questionnaire, where additional questions were asked in the second collection. processing

The informants

Encouraged by articles in the Swiss press, the informants sent some or all of the SMS they wrote during this time to a designated Swisscom number. Some informants just forwarded some or even every SMS they wrote, others seem to have added the Swisscom number as a default number to be added to every SMS. Two specific numbers were offered, one for the German and one for the Latin parts of Switzerland in the first call and one for the Italian and Romansh part respectively in the second call. However, these differentiations between regions was only used to communicate with the participants (i.e. to send the link to the questionnaire) in the appropriate language. They have no importance for the corpus whatsoever, because the language tagging that was performed later on improved upon a differentiation between languages that might have been deduced from these telephone-numbers. After sending us a first START-SMS, which was not included in the corpus, the participant received a link to the questionnaire to be filled in.

Sponsoring

The collection of the SMS contained in this corpus would not have been possible without Swisscom. Not only did they collect the individual SMS and personal data for us, they also ensured anonymity of the informants and thereby encouraged people to actually participate.

Privacy

No member of the team ever saw a phone number of the informants. People and their SMS can therefore not be traced back. Furthermore, the first step after the data collection was to remove any type of personal information from the corpus. These steps were performed by means of computational linguistics. They show a reliability of more than 90% so data can be assumed to comply with Swiss and international regulations about data privacy. If you still recognize authors of specific SMS based on the topics that they write about, you are asked to comply with common research ethics and keep that knowledge to yourself.