Cleaning the data up

No kind of censorship whatsoever was applied to the data. But still some SMS had to be removed from the original data, especially duplicate SMS that were created due to technical problems as well as SMS that were obviously not written by humans, such as notifications about new MMS etc.

Original request made to potential informants

The original request, that was published in the press, requested for SMS that were:

- Written by the very person who sent the SMS in. I.e. we did not want people to send in SMS that they had received but only the ones they had written to be sent out.
- Written on a mobile phone, i.e. we did not want to receive SMS that were composed on a computer and then sent via the internet.

These guidelines were not always followed by the people who sent us SMS. We did receive SMS that were composed on the computer. In some cases this was obvious because of the additional text "sent by Xtrazone" or similar, in other cases we are not aware of it. We also received SMS which were sent in by the recipient. Again, this is obvious in some cases, especially when several SMS were being copied together and sent in as one SMS. In this case, a short communication can be found in one single SMS. There are certainly other cases of recipients sending in SMS as well, but those cannot be spotted.

SMS removed

As said before, no SMS were removed based on their content. Even SMS which clearly disregarded our original instruction, i.e. SMS that were sent from the computer or contained received SMS were being kept. However, some 2'000 SMS were removed because they fulfilled one of the following condition:

1. Duplicate SMS: For technical reasons, we received two copies of some SMS that the users sent only once. Thus, if a couplet of SMS featured exactly the same text and exactly the same time stamp, one of them was deleted. There are also SMS in the corpus with exactly or nearly the same text as another one but with different time stamps. We did not consider these to be duplicates owed to technical reasons and thus did not remove them. The reasons for not removing them are the following. On the one hand, they were sent to us by people who participated in our campaign and we would have shown a lack of respect if we had deleted them. It can be assumed that these identical or nearly identical SMS were also sent to the intended recipient, thus, they are real data which is valuable to us. On the other hand, it would have been difficult to actually draw a line. If one full stop is replaced by an exclamation point in the second SMS, is it still the same one? And what if a typing error is corrected in the second SMS while the rest stays the same, is that another SMS or the same? We did not want to make decisions on those questions and thus valued and kept all the data that were being sent to us.
2. SMS, that were obviously not sent by human beings but by computers: we found some SMS, that have to be considered as automated SMS that were not sent but received by the people who sent in SMS. Among those are SMS that inform about an incoming MMS or SMS that were created by digital agendas. In the latter, we found texts like "9am: dentist. 10am: meeting with Bob". Another type of automated SMS were those that informed about general events, such as "tomorrow: special waste collection in your district". All these SMS do not feature the type of language we are interested in, i.e. a language that is used in SMS to communicate between human beings. We thus decided to delete them.

All other SMS were kept exactly as we received them except for anonymization.