

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 for chats that had to be [removed](#). Furthermore, [demographic data](#) (were provided) were linked to the chats.

In a first step the most basic processing of the data took place such as to allow the project members to work with the data. This included the [anonymization](#) and the annotation of a [main language](#) per chat and thus the creation of [subcorpora](#).

In a later step, more computational tasks were applied to the corpus. This included a more profound annotation of [languages](#) (i.e. each message was annotated for its language as opposed to the chat annotation performed in the first step), [part of speech annotations](#) were applied and the German dialectal data was normalized.

From:

<https://whatsup.linguistik.uzh.ch/> -

Permanent link:

https://whatsup.linguistik.uzh.ch/corpus/00_corpus?rev=1569177207

Last update: **2022/06/27 09:21**

