

# 1.1 Sub-corpora

Based on the annotation of the languages per chat, different sub-corpora were created.

The following basic considerations were applied when creating the sub-corpora:

## Definitions for sub-corpora

- Each chat was to be assigned to only one language-sub-corpus.
- Additionally, we differentiate between chats where we have demographic information for all participants and those where we do not. In the former case, the sub-corpus gets the extension \_DEMOG.
- Where additional tasks were performed on individual chats (e.g. normalization or part-of-speech tagging) we created additional sub-corpora per language.

## Main sub-corpora

- WUS: All data, i.e. the whole corpus
- WUS\_DEU: All data where non-dialectal German provides the most messages
- WUS\_DEU\_DEMOG: A subgroup thereof where we have demographic information from all communication partners.
- WUS\_FRA: All data where French provides the most messages
- WUS\_FRA\_DEMOG: A subgroup thereof where we have demographic information from all communication partners.
- WUS\_GSW: All data where dialectal German provides the most messages
- WUS\_GSW\_DEMOG: A subgroup thereof where we have demographic information from all communication partners.
- WUS\_ITA: All data where Italian provides the most messages
- WUS\_ITA\_DEMOG: A subgroup thereof where we have demographic information from all communication partners.
- WUS\_ROH: All data where Romansh provides the most messages
- WUS\_ROH\_DEMOG: A subgroup thereof where we have demographic information from all communication partners.

Additionally to these corpora, you also see corpora with lowercase letters in the browser (e.g. deu-rtagged, ita-tagged, roh etc.). These corpora contain data from our [SMS project](#).

## Smaller corpora

Next to these main sub-corpora, there are some smaller sub-corpora:

- WUS\_SMALL: Chats that are either smaller than 100 messages or where the majority of messages are not in a national language.
- WUS\_SMALL\_DEMOG: A subgroup thereof where we have demographic information from all communication partners.
- WUSdemographics: Only demographic data per person. This sub-corpus is much faster if you want to look up demographic data only.
- WUS\_ARGDROP and WUS\_ARGDROP\_language: Sub-corpora for which argument drop has been manually annotated. For the architecture of the annotations and scientific considerations behind it see [Stuntebeck, Franziska \(2018\): "Annotating Argument Drop in the Swiss WhatsApp](#)

Corpus". In: Generative Grammar in Geneva (GG@G) XI, 175-187.

## Other corpora in the browsing tool

Additionally to these corpora, you also see corpora with lowercase letters in the browser (e.g. deu-rftagged, ita-tagged, roh etc.). These corpora contain data from our [SMS project](#).

## More information about the subcorpora

The individual sub-corpora are well documented in terms of size etc. within the browsing tool. Check the according [section](#) for more information.

From:

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

Permanent link:

[https://whatsup.linguistik.uzh.ch/01\\_corpus/01\\_subcorpora?rev=1641390262](https://whatsup.linguistik.uzh.ch/01_corpus/01_subcorpora?rev=1641390262)

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

