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.

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.

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:
http://pub.linguistik.uzh.ch/whatsup-doc/ - **Documentation**

Permanent link:
http://pub.linguistik.uzh.ch/whatsup-doc/doku.php?id=01_corpus:01_subcorpora&rev=1588592542

Last update: **2020/05/11 08:55**