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. As mentioned in the section languages per chat, each chat has an annotation for languages that occur in 100 and more messages (meta::lang_100_and_more). However, there are chats with e.g. more than 100 messages in French and in German. In that case, the language that provides most chats is used for the chat to be assigned to a language-sub-corpus. Example: if a chat is built up of 150 messages in French and 120 messages in German, it appears in the main corpus (WUS) as well as in the French sub-corpora (WUS_FRA and WUS_FRA_DEMOG) but not in any of the German sub-corpora.
- 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 exist 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:
https://whatsup.linguistik.uzh.ch/ -

Permanent link:
https://whatsup.linguistik.uzh.ch/01_corpus/01_subcorpora?rev=1586932624

Last update: **2022/06/27 07:21**