1.3.4 Language per chat

In order to assign a language tagging to each chat, student helpers read through the first 250 messages and assigned two possible attributes per language:

- lang_100_and_more

Available languages:

- fra: French
- ita: Italian
- roh: Any variety of Romansh
- gsw: dialectal German as used in Switzerland
- deu: non-dialectal German
- eng: English
- spa: Spanish
- sla: Any Slavic language

Please note: In the browsing tool ANNIS, we created sub-corpora per language, where each message appears in one and only one sub-corpus, even though there may be several languages annotated as lang_100_and_more for a specific chat. In order to assign a chat to a sub-corpus, we arbitrarily prioritized the languages: ROH > GSW > FRA > DEU > ITA > ENG/SPA/SLA; Thus, e.g. a chat with English and German annotated with lang_100_and_more languages was assigned to the German subcorpus; a chat with the annotation lang_100_and_more for Romansh and and any other languages annotated as lang_100_and_more will only be found in the WUS_ROH(_DEMOG) subcorpus.

We thus considered the main language (as defined above) that provided the most messages to be the top main language and assigned it to the according subcorpus. If you want to work with all chats that contain a specific language in more than 100 messages, use the query *msg & meta::lang_100_and_more="fra, gsw"* on the whole corpus.

For an overview over languages in the corpus consult: Ueberwasser, Simone; Stark, Elisabeth (2017)2017: "What's up, Switzerland? A corpus-based research project in a multilingual country". In: Linguistik online, 84/5, 105-126. https://bop.unibe.ch/linguistik-online/article/view/3849/5834