1.2.4 Languages and varieties

Languages and varieties per chat

In order to assign a language tagging to each chat, we looked at the first 250 messages and assigned two possible attributes per language:

- lang_100_and_more: Languages that were found in more than 100 messages
- lang_less_than_100: Languages that were less frequent

for the following 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. In most cases, this it the language that delivers more than 100 chats. If there are two languages providing more than 100 messages, we arbitrarily prioritized the languages: ROH > GSW > FRA > DEU > ITA > ENG/SPA/SLA.

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 and varieties in the corpus consult: Ueberwasser, Simone; Stark, Elisabeth (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

Languages and varieties per message

The information of the main language of a message is saved in the annotation *most_likely_lang* and can thus be queried with e.g. `most_likely_lang="gsw"`.

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

Romansh varieties:

- roh-ja: Jauer Romansh
- roh-sr: romontsch sursilvan
- roh-st: rumàntsch sutsilvan
- roh-sm: rumantsch surmiran
- roh-pt: rumauntsch puter
- roh-vl: rumantsch vallader
- roh-gr: rumantsch grischun