

## 1.3.4 Languages and varieties

### Languages and varieties per chat

In order to assign a language tagging to each chat, we looked 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 is 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)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>

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

Permanent link:  
[https://whatsup.linguistik.uzh.ch/01\\_corpus/02\\_preprocessing/04\\_languages?rev=1587033122](https://whatsup.linguistik.uzh.ch/01_corpus/02_preprocessing/04_languages?rev=1587033122)

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

