1.2.4 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 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)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

1.3.5 Languages and varieties per message

In an iterative computational linguistic procedure based on n-grams, the most likely language per message was determined. In other words, the computational linguist looked for patterns of characters that are typical for a specific language/variant and then assigned this language/variant to all the words that showed this pattern. By comparing these annotations for tokens over the whole message, one language/variant was the "winner" and thus annotated as the most likely language/variant.

As an example, the characters <iich> are not likely to be found in any language or variant in the corpus except Swiss German dialect. If many such patterns appear in the same message, we take the most likely language variant to be Swiss German dialect. If more patterns identified as French appear, the most likely language is French etc.

The information extracted in this way 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

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

Permanent link:
https://whatsup.linguistik.uzh.ch/01_corpus/02_preprocessing/04_languages?rev=1587037158

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