Language per chat

In a first step and in order to provide adequate data to the researchers in the team, student helpers looked at every chat. Their approach was:

- Read through the individual chat until you have come across 100 messages in one and the same language (or variety in the case of German, where we differentiate between the Swiss German Dialect and not dialectal German).
- If at this point, any other language/variety provides more than 50 messages, read on until you have read a total of 250 messages. If, in the process of this reading, another language comes to 100 messages, consider both languages as main language. If not, only the language providing more than 100 messages is the main language.
- Mark the chat as containing one or more main languages (e.g. attribute: lang_100_and_more="fra, gsw"), also as lang_less_than_100="deu, eng, gsw" e.g. when only a few messages of e.g. these 3 languages are contained in the same chat.
- Take note of the other languages found in the course of this process (e.g. attribute: contains_eng="true").

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

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

Permanent link:
https://whatsup.linguistik.uzh.ch/01_corpus/02_preprocessing/04_languages?rev=1586891082

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