Sub-corpora

The corpus all-tagged contains all SMS in all languages. Data for all languages except Romansh are tagged with TreeTagger.

Next to that, the following sub-corpora per language are available:

- deu-rftagged: non-dialectal German data tagged with RF-Tagger
- deu-tagged: non-dialectal German data tagged with TreeTagger
- fra-tagged: French data tagged with TreeTagger
- gsw-rftagged: Swiss German data where the normalized data was tagged with RF-Tagger
- gsw-tagged: Swiss German data where the normalized data was tagged with TreeTagger
- ita-tagged: Italian data taggend with TreeTagger
- roh: Romansh data

The list of sub-corpora is also a good starting point to get information about available fields for your query, to get examples and statistics.

Please keep in mind that you also see corpora with upper-case letters in the browser (e.g. deu-rftagged, ita-tagged, roh etc.). These corpora contain data from our WhatsApp project.

Tokens and messages per sub-corpus

Next to the name of each sub-corpus, you see the number of SMS (marked as "Texts") and tokens. You can use these figures for statistics.

Information about the (sub-)corpora

When you press on the small `i` for information to the right of each (sub-)corpus name, you find more information about the corpus. More specifically:

- Some statistic information about the sub-corpus including an ULR pointing to this sub-corpus at the bottom.
- Information about the version to be quoted in publications.
- If you need specific information about an individual chat, you can select the SMS instead of the sub-corpus in the top left to get information such as languages contained, demographic information, etc. This is also an easy way to see which SMS are integrated in this sub-corpus.

On the right-hand side of the information window, you see which annotations are available to be queried for the selected sub-corpus.

- You have two categories of information: Node Annotations are attributes on token level. Meta Annotations contain information about the informant.
- To the right of the name of the annotation, there is an example query for that specific annotation. If you click on that text, a sample query is entered into the query field in the main screen. This is the easiest way to generate queries, since you can always modify it in the query

field. Example: if you click on mt_fr (Mother tongue French), an example like `node &
meta::mt_fr="false"` is entered into the query field. More precisely: `node` will fetch also all tokens that are in such SMS; if you want to distinguish between messages and tokens, you should explicitly query for one or the other: `tok & …` or `msg & ….`

List of chats in the sub-corpus

By clicking on the little piece of paper next to the information `i` in the list of sub-corpora, you get a list of all SMS in the respective sub-corpus.

From here, you can click on `full text` to view the whole SMS (without any annotations).

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

Permanent link:
https://sms.linguistik.uzh.ch/02_browsing/01_sub_corpora?rev=1643186241

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