2.1 Sub-corpora

As explained in the section about the creation of the sub-corpora, you can work with either the full corpus WUS or you can select different sub-corpora that mainly depend on the main language within the chat. You find the list of sub-corpora in the bottom left in ANNIS.

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.

Tokens and messages per sub-corpus

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

Please watch out: If you work with corpora where not all participants gave their permission to use their texts, the figure for tokens is off because messages without permission were replaced by texts like *redactedQ12tokens55characters* . These texts count as tokens, too. If you need statistics that depend on the number of tokens in a (sub-)corpus, you are advised to work with corpora with the extension _DEMOG.

Information about the corpus

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

- Some statistic information about the corpus including a link at the bottom for reference
- Information about the version to be quoted in publications.
- If you need specific information about an individual chat, you can select the chat instead of the corpus in the top left to get information such as number of messages, number of speakers, etc. This is also an easy way to see which chats are integrated in this sub-corpus.

Still in the information windows, you see which annotations are available to be queried for this specific sub-corpus on the right.

- You have two categories of information: **Node Annotations** is information that was provided by the informant or that was added by us (e.g. part of speech annotations), while **Meta Annotations** contains information that was generated automatically (e.g. number of messages in a chat or languages).
- 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 Lang_100_and_more, an example like *node & meta::lang_100_and_more="fra, eng*" is entered into the query field. This query would search for messages in chats with more than 100 messages in French and in English. In the query field, you can not replace the languages to search for chats with more than 100 messages in Italien by replacing *fra, eng* by *ita*.

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 chats in this specific sub-corpus.

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

Permanent link:
https://whatsup.linguistik.uzh.ch/02_browsing/01_sub_corpora?rev=1573114564

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