

2.1 Sub-corpora

As explained in [section 1.1](#), you can work with either the full corpus WUS or you can select different sub-corpora. 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.

Please keep in mind that you also see corpora with lowercase letters in the browser (e.g. deu-ftagged, ita-tagged, roh etc.). These corpora contain data from our [SMS project](#).

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 note: If you work with corpora where not all participants gave their [permission](#) to use their messages, the figure for tokens is off because messages without permission were replaced by messages 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 (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 chat instead of the sub-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.

Corpus information for WUS_FRA_DEMOG (ID: 51867)

Metadata		Available annotations	
Select corpus/document: WUS_FRA_DEMOG		Node Annotations	
Name	Value	Edge Annotations	
editors	Anne Göhring, Beni Ruerf, Simone Lieberwasser	Edge Types	
longName	whatsUpSwitzerland FRA demographics subcorpus	Meta Annotations	
project	whatsUpSwitzerland	Name	Example (click to use query)
releaseDate	20190815	total_msg	node & meta:total_msg="14789"
shortName	WUS_FRA_DEMOG	consent_speakers	node & meta:consent_speakers="2"
version	7.0	contains_eng	node & meta:contains_eng="true"
		demographics	node & meta:demographics="2"
		doc	node & meta:doc="chat195"
		content_msg	node & meta:content_msg="6838"
		empty_msg	node & meta:empty_msg="0"
		media_msg	node & meta:media_msg="69"
		lang_less_than_100	node & meta:lang_less_than_100="eng"
		encrypted_msg	node & meta:encrypted_msg="0"
		lang_100_and_more	node & meta:lang_100_and_more="fra"
		speakers	node & meta:speakers="2"
		no_consent_msg	node & meta:no_consent_msg="0"

Link to corpus: <http://linguistik.sms.uzh.ch/9080/anne-goehring/cvYVYX025Qv9ERU1PRw>

Figure 1: Information about a (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.

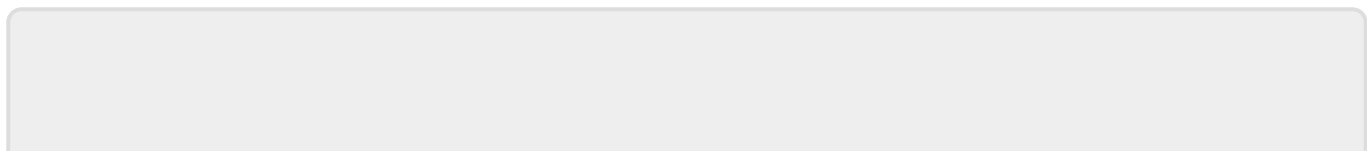
- You have two categories of information: Node Annotations are attributes on either token and message level. Meta Annotations contain information at the chat level; most of the meta annotations indicate sizes (e.g. total number of messages in a given chat) and were automatically computed.
- To the right 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. More precisely: node will fetch also all tokens that are in such chats; 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 chats in the respective sub-corpus.

From here, you can click on complete chat view to view the whole chat (without any annotations). Once in this list of messages, you can always click on an individual message ID to see that message with its annotations.

If you click on the little *i* at the very right of the list of chats, you see all the meta information about the respective chat.



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

Permanent link:
https://whatsup.linguistik.uzh.ch/02_browsing/01_sub_corpora?rev=1641390328

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

