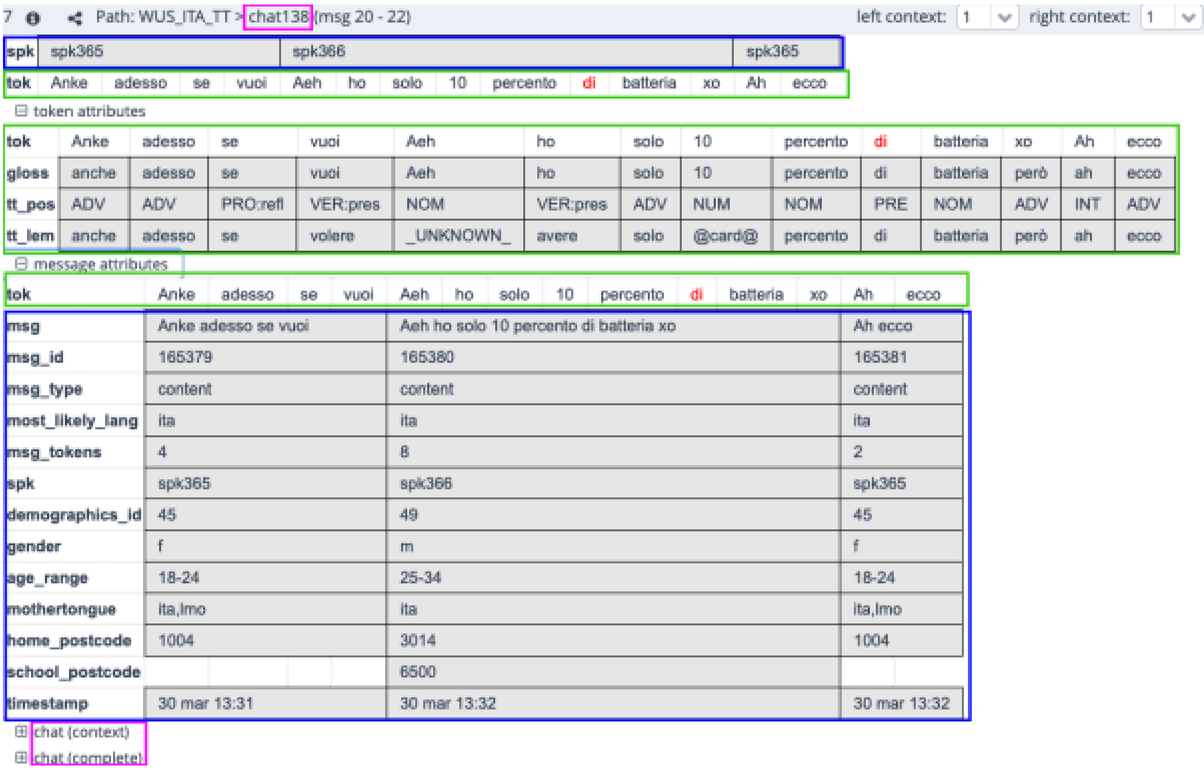
2.2 Layers of information

WhatsApp messages are built up in a hierarchy: a chat contains messages that contain tokens that contain characters. A corpus of WhatsApp chats should allow for all these layers to be queried. Additionally, there is meta-data about the chats (e.g. number of messages) and about the messages (e.g. the timestamp when it was written) and about the informant (e.g. his/her age) and about the tokens (e.g. part of speech). This makes our corpus a rather challenging and complex endeavor.

These layers can nicely seen when browsing results from a query:



7 Path: WUS_ITA_TT > chat138 (msg 20 - 22) left context: 1 right context: 1

spk	spk365				spk366								spk365	
tok	Anke	adesso	se	vuoi	Aeh	ho	solo	10	percento	di	batteria	xo	Ah	ecco

⊟ token attributes

tok	Anke	adesso	se	vuoi	Aeh	ho	solo	10	percento	di	batteria	xo	Ah	ecco
gloss	anche	adesso	se	vuoi	Aeh	ho	solo	10	percento	di	batteria	però	ah	ecco
tt_pos	ADV	ADV	PRO:refl	VER:pres	NOM	VER:pres	ADV	NUM	NOM	PRE	NOM	ADV	INT	ADV
tt_lem	anche	adesso	se	volere	_UNKNOWN_	avere	solo	@card@	percento	di	batteria	però	ah	ecco

⊟ message attributes

tok	Anke	adesso	se	vuoi	Aeh	ho	solo	10	percento	di	batteria	xo	Ah	ecco
msg	Anke adesso se vuoi				Aeh ho solo 10 percento di batteria xo								Ah ecco	
msg_id	165379				165380								165381	
msg_type	content				content								content	
most_likely_lang	ita				ita								ita	
msg_tokens	4				8								2	
spk	spk365				spk366								spk365	
demographics_id	45				49								45	
gender	f				m								f	
age_range	18-24				25-34								18-24	
mothertongue	ita,lmo				ita								ita,lmo	
home_postcode	1004				3014								1004	
school_postcode					6500									
timestamp	30 mar 13:31				30 mar 13:32								30 mar 13:32	

⊞ chat (context)
⊞ chat (complete)

Chats

In this example, you find the chat back as an ID (chat138) at the top in pink. If you want to see the whole chat, you see two options at the very bottom: chat in context (faster) or the whole chat (can be slow). When you click on the little <i> in the top bar, you can also see meta data about the chat, such as the number of speakers, languages, total messages etc.

Messages

In this pink chat, you see three selected messages in blue:

- Message 165379: Anke adesso se vuoi
- Message 165380: Aeh ho solo 10 percento di batteria xo
- Message 165381: Ah ecco

As you can see, these messages have meta data assigned to them, as well, e.g. the message ID and the speaker (these pieces of information are always available) as well as information provided by the informant such as age, mothertongue etc.

Tokens

The individual tokens are annoted in green in the above example and they are aligned to the message, to which they belong.

Tokens, too, (can) have meta data that is assigned to them. In the example shown above, you have the following meta data that was created by our team or by our computational linguists:

- Gloss: a normalization, i.e. a "translation" into standard spelling. A good example here is *xo*, which was normalized as <però>.
- tt_pos: A part-of-speech annotation generated with the parser TreeTagger.
- tt_lem: The lemma for each token as it was created by TreeTagger.

The red token *di*, by the way, is the one that we queried for to create this screen shot.

Labels

On all three layers, i.e. for chats, messages and tokens, as well as for all the meta data, you see the labels, e.g. msg_id, gloss, home_postcode etc. These labels are used for queries.

Examples:

- If you want to see the whole message 165380, your query would be *msg_id="165380"*
- If you want to find verbs in the present tense, your query is *tt_pos="VER:pres"*

To see all the labels available in a specific sub-corpus, check the information for the sub-corpus.