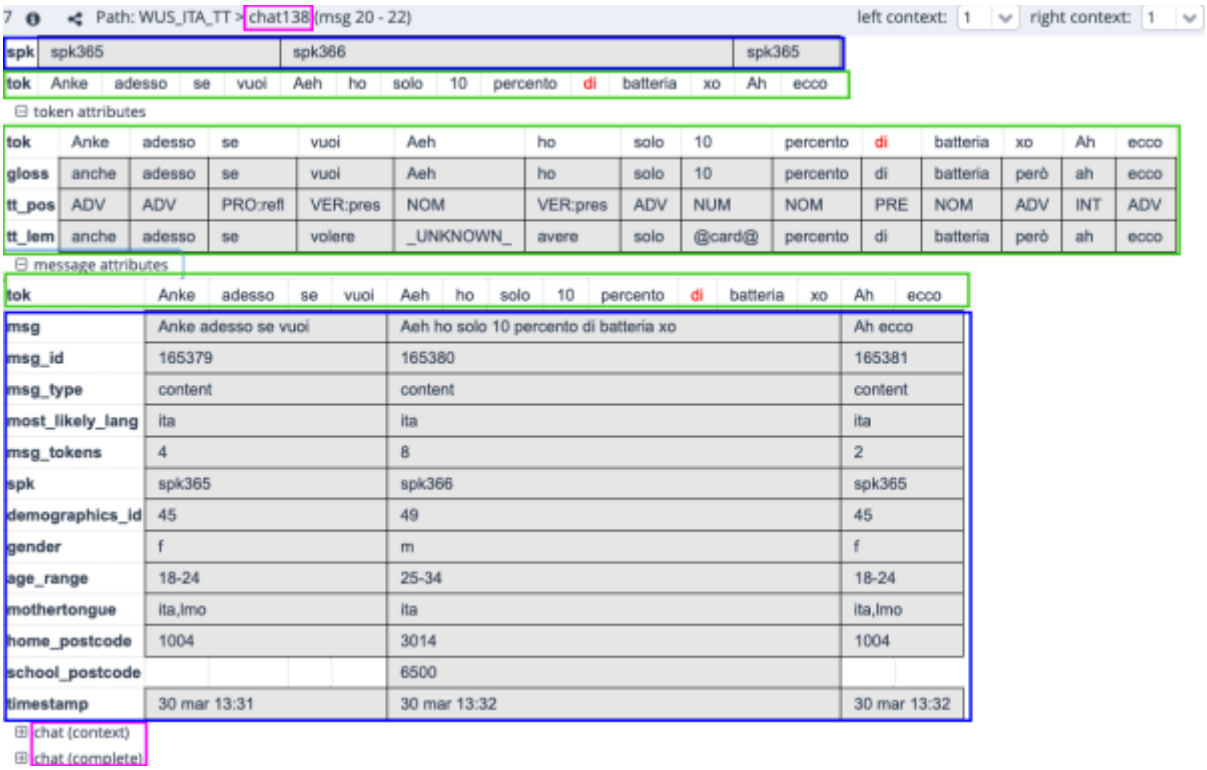
2.2 Layers of information

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:



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.