2.4 Queries

You can query the corpus for very different pieces of information such as messages written in the chats, part of speech annotations, demographic information like the age of the informant, or statistical information like the number of messages in a chat.

Please keep in mind that all the fields in the corpus are text fields. For your query that means that you cannot search for *larger* or *smaller than*. For example, you cannot say "show me all chats with more than 1000 messages", because this field is interpreted as text and not as a digit. In theory, you would have to query for messages with 1000, 1001, 1002, 1003 etc. messages. In practice, this is not a very useful criterion for a query.

The following three options for querying the corpus are described in more detail in the sub-sections of this document:

- Simple queries: these are basically queries for words e.g. *est* or *ich* etc.
- RegEx queries: are used for more complex patterns such as alternatives (*man* and *men*), for patterns with different endings (*Man* and *Manchester*) etc.
- Queries for meta data
- Combined queries: are used whenever you want information from different layers, e.g. the word *man* written by only females.

Please remember to always keep in mind the unit that you are querying. If you query in individual tokens, you do not have to consider separators such as spaces, punctuation, tabs etc. If, on the other hand, you work on a whole message, you have to take such things into account. You also have to remember that querying over whole messages is very slow and can lead to time outs.