Questions

Segmentation

- **How should you segment sentences with semi-colon? As a single sentence or as two sentences? Should it depend on context?**

Semi-colon usually means that the sentence is not over yet but it also separates different clauses in the sentence. So, if we want to work on the clauses level of a sentence and not the entire sentence, we should segment sentences with semi-colon into two sentences (or clauses in this case).

- **Should sentences with ellipsis... be treated as a single sentence or as several sentences?**

Sentences with ellipsis end with a trail of dots (usually 3 consecutive dots). In some cases, ellipsis can occur in the middle of the sentence as well. Therefore, they should be treated as a single token and/or sentence and not multiple sentences.

- **If there is an exclamation after the first word in the sentence should it be a separate sentence? How about if there is a comma?**

If there is an exclamation mark after the first word, it should be treated as a single token and not a separate sentence. Comma does not indicate a separate sentence.

- **Can you think of some hard tasks for the segmenter?**

When there are abbreviations, it is difficult for segmenter to separate sentences.

Tokenization

- **Why should we split punctuation from the token it goes with ?**

Punctuation is not part of the word in most cases and should therefore be treated as a separate token.

- **Should abbreviations with space in them be written as a single token or two tokens ?**
  - **How about numerals like 134 000 ?**

Abbreviations should be treated as one token and same with numerals.

- **If you have a case suffix following punctuation, how should it be tokenised ?**

For certain punctuations such as hyphen(-), case suffix should be treated as the same token as punctuation mark.

- **Should contractions and clitics be a single token or two (or more) tokens ?**

They should be taken as a single token.