Segmentation

1. Sentences with semicolons doesn’t have to be two different sentences. It depends on the context for segmentation.
2. sentences with ellipsis should generally be treated as a single sentence. However, there may be cases where an ellipsis is used to represent a series of incomplete thoughts, each of which could be treated as separate sentences. In such cases, it's important to consider the overall meaning and flow of the text when deciding how to segment sentences.
3. If there is an exclamation after the first word in a sentence, it should typically be considered as part of the same sentence. The presence of an exclamation mark does not indicate the start of a new sentence. Similarly, a comma after the first word may or may not indicate a new sentence depending on its context within the sentence.
4. Segmentation may be difficult in sentences with informal text where punctuation may be used less strictly.

Tokenization

1. There might be some ambiguity if punctuations are not separated. Moreover, splitting punctuation helps in understanding the semantic structure more clearly.
2. Choice of separating abbreviations with spaces completely depends on the context and particular task. For instance, if the abbreviation with spaces is considered a single unit of meaning or if preserving the space is important for the context. When it comes to numerals like 134 000, If the space between is separated then the value changes. Hence the space should be preserved in such cases.
3. If the punctuation and the case suffix are distinct entities, they should be tokenized separately. This also helps capturing the punctuations separately.
4. Contractions and clitics are shorter forms which are created by combining two words. Hence, they should not be tokenized and should be treated as one word.