01 Tokenization

Tokenization is the process of dividing text into individual tokens (words, digits, punctuation). For the machine annotation of corpora in the Slovenian context, we currently use the CLASSLA-Stanza tagger, more precisely the Obeliks tokeniser included in it. The rules guiding the automatic tagging are also adhered to during manual revision.

Annotation Guidelines

This chapter summarizes the annotation guidelines for tokenization.

 Space is the principal separator for tokens.

 Sequences of words that can be written both with or without space without changing its meaning (e.g. **kdorkoli**, **kdor koli** “anybody, any body”) follow the same principle and become either one or two tokens depending on the use of space.

 During tokenization, all characters are divided into two categories: words (W) and characters (C).

 C tokens are recognized on the basis of a predefined list of punctuation- and symbol-like characters included in the tokenizer (depending on the annotation system, e.g. Universal Dependencies or JOS/MULTEXT-East) and consist of single characters only. Sequences of two or more characters (e.g. **?!**) are treated as sequences of separate C tokens.

 If a string of alphanumeric characters between two spaces includes C characters, it is usually split into several tokens (e.g. **AC/DC** and **Micro$oft** are split into three tokens 'AC' '/' 'DC' and 'Micro' '$' 'oft').

 However, the following exceptions, in which C characters become parts of W tokens, apply:

  Apostrophe becomes part of a W token if used without space on both sides (e.g. **O’Brian** "O’Brian", **mor’va** "we have to").

  Comma and colon become part of a W token if used without space on both sides and if the string contains only digits (e.g. **30:00**, **200,000,000**).

  Hyphen becomes part of a W token if used without space on both sides and if:

- the left part is an acronym (in capital letters), a single letter or a digit

- the right part is an affix or an inflectional ending; a finite list of possible affixes and endings is integrated in the tokenizer, e.g. **OZN-ovski** "similar to United Nations", **a-ju** "to the letter a", **15-i** "the 15th".

  Dot becomes part of a W token if it is:

- used without space on both sides and the string contains only digits, e.g. **1.2**

- used without space on the left and is part of an abbreviation or ordinal number (e.g. **dr.**, **4.**, **IV.**); a finite list of possible abbreviations is integrated in the tokenizer.

 All C characters become part of a single W token in strings recognized as URLs or addresses using a regular expression.

Information on whether a token is not followed by a space (e.g. **d.o.o.** vs. **d. o. o.**) is indicated with SpaceAfter=No feature in the MISC column.

References and Links

This chapter compiles relevant references and provides links to projects where the lemmatization process has been developed and applied to Slovene texts.

Projects, in which the system has been developed or applied

Universal Dependencies

MULTEXT-East - Multilingual corpora and text tools for Central and East European langauges

JOS - Linguistic Annotation of Slovene: Methods and Resources

Communication in Slovene

Janes - Resources, Tools and Methods for the Research of Nonstandard Internet Slovene

Development of Slovene in a Digital Environment

The Obeliks tool for tokenization and sentence segmentation

https://github.com/clarinsi/obeliks

References

Krek, Simon; Erjavec, Tomaž; Dobrovoljc, Kaja; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka in Brank, Janez. The ssj500k training corpus for Slovene language processing. V: Fišer, D. in Erjavec, T. Jezikovne tehnologije in digitalna humanistika: zbornik konference: 24.-25. september 2020, Ljubljana, Slovenija. Ljubljana: Inštitut za novejšo zgodovino, 2020. Str. 24–33.

http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_Krek-et-al_The-ssj500k-Training-Corpus-for-Slovene-Language-Processing.pdf [PDF]