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.

Revision #12
Created 13 November 2023 12:50:23 by Tina Munda
Updated 30 November 2023 12:03:35 by Tina Munda