

# 02 Segmentation

Segmentation is the process of dividing text into individual sentences. For the machine annotation of corpora in the Slovenian context, we currently use the CLASSLA-Stanza tagger, more precisely the Obeliks segmentator included in it. The rules guiding the automatic tagging are also adhered to during manual revision.

- [Introduction to Segmentation](#)
- [Annotation Guidelines](#)
- [References and Links](#)

# Introduction to Segmentation

This chapter summarizes the annotation guidelines for sentence segmentation.

The main guideline for demarcating sentences is a combination of final punctuation, space, and a capitalized word. This is supplemented with additional rules that cover abbreviations. These are written with a period, which can also serve as final punctuation (when the abbreviation is at the end of a sentence, e.g., 'itd.') or not (when the abbreviation is in the middle of a sentence, for instance 'itj.'). The final list of abbreviations that fall into either category is included in the Obeliks tool.

For segmenting Slovene non-standard texts, additional rules apply:

- In an entire tweet, check whether automatic sentence segmentation is correct. In the guidelines, the end of a sentence is marked for easier understanding with the symbol ¶.
- If part of the tweet functions as an independent sentence, it is treated as such (“@multikultivator Najbrž ne . ¶ :) ¶ Kot rečeno : bolje BO . ¶ Zrihtamo , ko utegnemo . ¶ ( PS : tudi v veselje " konkurence " ;)").
- The criterion for the end of a sentence is mainly a punctuation mark that acts as the final one in a sentence, e.g., period, exclamation point, question mark, quotation marks, or ellipsis (“Kaj praviš ? ¶ Aha !”).
- Unless there's a good reason to treat something as two sentences, it should remain one (“@urosgruber pri meni naloži CSS .. kar pa ne pomeni , da stran zgleda lepo :)” → one sentence because the dots acts more like a comma than a period).
- The end of a tweet is automatically also the end of a sentence, so this is not marked.

Complex cases:

□ Three dots:

□ Ponavadi je končno ločilo (“@SLO\_Super\_Visor po moje se jo izogiba kot hudič križa. ¶ Glavn da on spet laja ... ¶ :-))))))”).

□ Sometimes three dots indicate just an ellipsis or a pause in the middle of a sentence – in such a case, it's not final punctuation (“To se mi zdi ... neumno.”).

□ Names (@name), emoticons (\o/) or emojis (😊), and hashtags (#hashtag):

□ If they appear in the middle of a sentence, they are part of the sentence (“neka baka :) uleti pa praša če loh gre kr naprej”, “sej #tarca je pa dons kr ok”, “sej je rekla @Sandra d je treba to drgac”).

□ If they appear at the beginning of a tweet, they are treated as part of the first sentence (“@TadejTrcekTITO @lucijausaj @JJansaSDS titek, ne seri. odv. častno razsodišče je JE zgolj za odvetnike.”).

□ If they replace the final punctuation, they mark the end of a sentence (“kot da je to važn :)” ¶ nobenga to ne briga vec sploh”).

□ If they follow the final punctuation, they are treated as an independent sentence (“Sonce, sneg in pot pod noge! ¶ :) ¶ Gremo v hribe!”).

□ If several names, emoticons, or hashtags are strung at the end of a sentence, the last element is considered the end of the sentence (“itak ne morm sploh keša dvignt :) @tibonalta #broke” → the end of the sentence is the hashtag #broke).

Segmentation of spoken Slovene is currently done manually based on prosodically or semantically completed units.

# Annotation Guidelines

This chapter summarizes the annotation guidelines for segmentation as applied to Slovene texts.

## **Version 1.0 for non-standard Slovene**

**Project [Janes - Resources, Tools and Methods for the Research of Nonstandard Internet Slovene](#)**

ČIBEJ, Jaka, ARHAR HOLDT, Špela, ERJAVEC, Tomaž, FIŠER, Darja, ZUPAN, Katja, 2016: *Smernice za označevanje računalniško posredovane komunikacije: tokenizacija, stavčna segmentacija, normalizacija, lematizacija in oblikoskladenjsko označevanje*. Različica 1.0. Rezultat projekta Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine. [\[PDF\]](#) - only in Slovene

# References and Links

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

## **Projects, in which the system has been developed or applied**

[JOS - Linguistic Annotation of Slovene: Methods and Resources](#)

[Communication in Slovene](#)

[Universal Dependencies](#)

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