

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 languages](#)

[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]

Revision #4

Created 14 November 2023 09:56:54 by Tina Munda

Updated 5 December 2023 14:50:10 by Tina Munda