05 Lemmatization

When tagging text, each word form is assigned a lemma (the base form of the word), facilitating further processing in a unified way. The lemmatization system was developed in the project JOS: Linguistic Annotation of Slovene (Holozan et al. 2008) and follows the MULTEXT-East v4 or JOS morphosyntactic system in determining parts of speech, capitalization and some other features.

- Annotation Guidelines
- References and Links

Annotation Guidelines

This chapter summarizes the annotation guidelines for the lemmatization of Slovene texts. The guidelines are arranged from the latest, up-to-date version to the oldest version.

Version 2.0 (25-02-2023)

Project Development of Slovene in a Digital Environment
HOLOZAN, Peter, KREK, Simon, PIVEC, Matej, RIGAČ, Simon, ROZMAN, Simon, VELUŠČEK, Aleš, PORI, Eva, ARHAR HOLDT, Špela, 2023: *Specifikacije za učni korpus: lematizacija in MSD.* Različica 2.0. Rezultat projekta Razvoj slovenščine v digitalnem okolju.

[DOCX] [PDF] - only in Slovene

Version 1.0 (2008)

Project Communication in Slovene
HOLOZAN, Peter, KREK, Simon, PIVEC, Matej, RIGAČ, Simon, ROZMAN, Simon, VELUŠČEK, Aleš, 2008: *Specifikacije za učni korpus.* Različica 1.0. Kazalnik K2 projekta Sporazumevanje v slovenskem jeziku. [PDF] - only in Slovene

Version 1.0 for non-standard Slovene (21-12-2016)

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 the lemmatization of Slovene has been developed and applied to Slovene texts.

Projects, in which the system has been developed:

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

Training corpora containing manually revised lemmas:

Arhar Holdt, Špela; Krek, Simon; Dobrovoljc, Kaja; Erjavec, Tomaž; Gantar, Polona; Čibej, Jaka; Pori, Eva; Terčon, Luka; Munda, Tina; Žitnik, Slavko; Robida, Nejc; Blagus, Neli; Može, Sara; Ledinek, Nina; Holz, Nanika; Zupan, Katja; Kuzman, Taja; Kavčič, Teja; Škrjanec, Iza; Marko, Dafne; Jezeršek, Lucija; Zajc, Anja, 2022, Training corpus SUK 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1747.

Lenardič, Jakob; Čibej, Jaka; Arhar Holdt, Špela; Erjavec, Tomaž; Fišer, Darja; Ljubešić, Nikola; Zupan, Katja; Dobrovoljc, Kaja, 2022, CMC training corpus Janes-Tag 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1732.

References:

Erjavec, Tomaž; Fišer, Darja; Krek, Simon in Ledinek, Nina. 2010. The JOS Linguistically Tagged Corpus of Slovene. V: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valeta, Malta, Maj. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf [PDF]

Erjavec, Tomaž. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. Language Resources and Evaluation, 46(1): 131–142. DOI: 10.1007/s10579-011-9174-8 [PDF]

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]