03 Normalization

Computer-mediated communication (CMC) language significantly diverges from the standard language, posing challenges for current automatic text annotation tools. Normalization is essential for enhancing further text processing because it provides a standard equivalent for each non-standard occurrence. This step is critical as both lemmatization and morphosyntactic annotation of CMC language rely on these normalized forms (Čibej et al. 2016).

- Introduction to Normalization
- Annotation Guidelines
- References and Links

Introduction to Normalization

This chapter summarizes the process of normalizing non-standard Slovene words. A more detailed presentation can be found in the guidelines in the Annotation Guidelines chapter.

In the case of Slovene tweets, normalization was carried out simultaneously with tokenization, as shown in Table 1.
During the manual revision, five types of corrections were identified:

- A word requiring only a direct normalized form, e.g. **tukó** was manually normalized to **tako**.
- Words incorrectly written as one and requiring split into several tokens and, if neccessary, subsequent normalization, e.g. **nauta** (dialectal spelling of "you two will not" in Slovene) was manually normalized to **ne** ("not") and **bosta** ("you two will"), with the space between the two tokens indicated by a pipe | (see Table 1).
- A word incorrectly segmented into multiple tokens that needed merging and possibly further normalization, such as changing **o ga bn** to **ogabno**; redundant tokens were marked with $0 (see Table 1).
- A word erroneously split by the tokenizer, such as **s'm** (dialectal spelling of the word form 'sem' ("am") of the verb 'biti' "to be") being incorrectly tokenized as **s** + **'** + **m**, which were then manually merged and normalized as needed.
- Words erroneously merged by the tokenizer, for instance **5km** as one token, which had to be manually split and normalised as needed.

Token	Tokenizataion	Normalization
zato		
tukó		tako
nauta		ne l bosta
s	s’m	sem
‘	$0	$0
m	$0	$0
pršva		prišla

Table 1: Normalization and tokenization of a tweet.

Annotation Guidelines

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

Version 2.0

Project Development of Slovene in a Digital Environment

LENARDIČ, Jakob in FIŠER, Darja, 2022: *Smernice za ročno normalizacijo Janes Norm 3.0*. Rezultat projekta Razvoj slovenščine v digitalnem okolju. [DOCX] [PDF] - only in Slovene

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 the normalization process has been developed and applied to Slovene texts.

Projects, in which normalization has been developed or applied

Development of Slovene in a Digital Environment

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

Training corpora containing manually revised normalization

- **Janes-Tag:**

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.

Erjavec, Tomaž; et al., 2019, CMC training corpus Janes-Tag 2.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1238.

Erjavec, Tomaž; et al., 2017, CMC training corpus Janes-Tag 2.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1123.

Erjavec, Tomaž; Fišer, Darja; Čibej, Jaka; Arhar Holdt, Špela and Ljubešić, Nikola, 2016, CMC training corpus Janes-Tag 1.2, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1085.

Erjavec, Tomaž; Fišer, Darja; Čibej, Jaka and Arhar Holdt, Špela, 2016, CMC training corpus Janes-Tag 1.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1081.

Erjavec, Tomaž; Fišer, Darja; Čibej, Jaka and Arhar Holdt, Špela, 2016, CMC training corpus Janes-Tag 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1079.

- **Janes-Norm:**

Lenardič, Jakob; Čibej, Jaka; Arhar Holdt, Špela; Erjavec, Tomaž and Fišer, Darja, 2022, CMC training corpus Janes-Norm 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1733.

Erjavec, Tomaž; Fišer, Darja; Čibej, Jaka and Arhar Holdt, Špela, 2016, CMC training corpus Janes-Norm 1.2, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042,

http://hdl.handle.net/11356/1084.

Erjavec, Tomaž; Fišer, Darja; Čibej, Jaka and Arhar Holdt, Špela, 2016, CMC training corpus Janes-Norm 1.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042,

http://hdl.handle.net/11356/1083.

- **Janes-Syn:**

Arhar Holdt, Špela; Erjavec, Tomaž and Fišer, Darja, 2017, CMC training corpus Janes-Syn 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042,

http://hdl.handle.net/11356/1086.

References

FIŠER, Darja, LJUBEŠIĆ, Nikola, ERJAVEC, Tomaž. 2020. The Janes project: language resources and tools for Slovene user generated content. Language Resources and Evaluation. DOI:

10.1007/s10579-018-9425-z

ČIBEJ, Jaka, ARHAR HOLDT, Špela, ERJAVEC, Tomaž, FIŠER, Darja. Ročno označeni korpusi Janes za učenje jezikovnotehnoloških orodij in jezikoslovne raziskave. V: FIŠER, Darja (ur.). Viri, orodja in metode za analizo spletne slovenščine. Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 2018. https://ebooks.uni-lj.si/zalozbaul//catalog/view/111/203/2416-1 [PDF] - only in Slovene

LJUBEŠIĆ, Nikola, ERJAVEC, Tomaž, FIŠER, Darja. Orodja za procesiranje nestandardne slovenščine. V: FIŠER, Darja (ur.). Viri, orodja in metode za analizo spletne slovenščine. 1. izd. Ljubljana: Znanstvena založba Filozofske fakultete, 2018. Str. 74-98, 381-382, tabele. Zbirka Prevodoslovje in uporabno jezikoslovje. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/111/203/2413-1 [PDF]. - only in Slovene

FIŠER, Darja (urednik). Viri, orodja in metode za analizo spletne slovenščine. 1. izd. Ljubljana: Znanstvena založba Filozofske fakultete, 2018. 396 str., ilustr. Zbirka Prevodoslovje in uporabno jezikoslovje. https://ebooks.uni-lj.si/zalozbaul//catalog/book/111[PDF] - only in Slovene

ČIBEJ, Jaka, ARHAR HOLDT, Špela, ERJAVEC, Tomaž, FIŠER Darja. Razvoj učne množice za izboljšano označevanje spletnih besedil. V: Zbornik konference Jezikovne tehnologije in digitalna humanistika 2016, 40–46. https://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Cibej-et-al_Razvoj-ucne-mnozice.pdf [PDF] - only in Slovene

ERJAVEC, Tomaž, ČIBEJ, Jaka, ARHAR HOLDT, Špela, LJUBEŠIĆ, Nikola, FIŠER, Darja. Gold-standard datasets for annotation of Slovene computer-mediated communication. In Proceedings of RASLAN 2016: Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2016, pp. 29-40. https://nlp.fi.muni.cz/raslan/raslan16.pdf [PDF]

ČIBEJ, Jaka, FIŠER, Darja, ERJAVEC, Tomaž. Normalisation, tokenisation and sentence segmentation of Slovene tweets. Proceedings of the Workshop on Normalisation and Analysis of Social Media

Texts (NormSoMe). 2016, pp. 5-10. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf [PDF]