03 Normalizacija

Jezik spletne komunikacije se v marsikaterem vidiku razlikuje od standardnega jezika. Obstoječa orodja za označevanje besedil se z njim težje spopadajo. Normalizacija, katere cilj je vsaki nestandardni pojavnici pripisati standardno ustreznico, je ključna za izboljševanje nadaljnje obdelave besedil, saj tako lematizacija kot tudi oblikoskladenjsko označevanje nestandardnega jezika potekata na podlagi normaliziranih oblik (Čibej et al. 2016).

- Predstavitev normalizacije
- Označevalne smernice
- Reference in povezave

Predstavitev normalizacije

V tem poglavju je strnjeno predstavljen potek normalizacije nestandardnih besed. Podrobnejšo predstavitev najdete v smernicah v poglavju Označevalne smernice.

Normalizacija tvitov, v tabeli razdeljenih na pojavnice, je potekala hkrati s tokenizacijo.
Pri ročnem pregledu je bilo odkritih 5 vrst popravkov:

- Beseda, ki ji je bilo treba zgolj pripisati normalizirano ustreznico, npr. **tukó**, ki je bil ročno normaliziran v **tako**.
- Več besed, ki so bile nestandardno zapisane skupaj in jih je bilo treba razdružiti ter po potrebi še normalizirati, npr. **nauta** v **ne** in **bosta**, pri čemer je presledek med pojavnicama označen z navpičnico | (gl. Tabelo 1).
- Beseda, ki je bila nestandardno zapisana v več pojavnicah in jo je bilo treba združiti v eno in združek po potrebi normalizirati, npr. **o ga bn** v **ogabno**; odvečne vrstice (gl. Tabelo 1) so bile označene s sosledjem znakov $0.
- Beseda, ki jo je tokenizator avtomatsko razdružil; npr. **s'm** (narečni zapis oblike »sem« pomožnega glagola) v tri pojavnice: **s** + **'** + **m**. Tovrstne pojavnice so bile najprej ročno združene in po potrebi normalizirane.
- Beseda, ki jo je tokenizator avtomatsko združil, npr. **5km** kot ena pojavnica, kar je bilo treba ročno razdružiti in po potrebi normalizirati.

Pojavnica	tokenizacija	normalizacija
zato		
tukó		tako
nauta		ne l bosta
s	s’m	sem
‘	$0	$0
m	$0	$0
pršva		prišla

Tabela 1: Normalizacija in tokenizacija tvita.

Označevalne smernice

V tem poglavju so zbrane označevalne smernice za normalizacijo nestandardnih besedil. Smernice so razvrščene od zadnje, ažurne različice do nastarejše različice.

Različica 2.0

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

Različica 1.0 za nestandardno slovenščino

projekt Janes - Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine
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]

Reference in povezave

V tem poglavju so zbrane relevantne reference in povezave na projekte, v katerih se je postopek normalizacije razvijal in uporabljal.

Projekti, na katerih se je razvijal označevalni sistem

Razvoj slovenščine v digitalnem okolju

Janes: Viri, orodja in metode za raziskovanje nestandardne spletne slovenščine

Učni korpus z ročno pregledano normalizacijo

- **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.

Reference

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]

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].

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]

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]

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]