04 Oblikoskladnja MULTEXT-East

Sistem MULTEXT-East za označevanje oblikoskladnje pojavnic v korpusih definira črkovne kode, npr. »Somei«, in njihovo preslikavo v slovnične lastnosti, npr. »samostalnik vrsta=občno_ime spol=moški število=ednina sklon=imenovalnik«. Sistem je definiran za 20 jezikov oz. dialektov, mdr. za vse slovanske jezike.
Uporaba oznak MULTEXT-East za slovenščino se je začela že leta 1996, nato pa se je nadaljevala pri vseh nadaljnjih ročno in tudi avtomatsko označenih odprtih korpusih slovenščine, vse do pojavitve sistema označevanja oblikoskladnje v okviru projekta Universal Dependencies, ki sedaj počasi prevzema vlogo, ki jo je desetletja imel MULTEXT-East.

- Predstavitev oznak
- Označevalne smernice
- Reference in povezave

Predstavitev oznak

V tem razdelku na kratko opišemo zasnovo specifikacij MULTEXT-East in podamo povezave do specifikacij.

Večjezične specifikacije MULTEXT-East so zapisane v XML, po priporočilih TEI, in definirajo oblikoskladenjske značilke (atribute in njihove vrednosti) besed, tj. značilnosti besed, ki so na preseku oblikoslovja in skladnje. Specifikacije podajo tudi preslikavo množic teh značilk v oblikoskladenjske opise (angl. *morphosyntactic descriptions*; MSD), ki so kompaktni nizi, uporabljeni pri označevanju korpusov. Tako se na primer MSD "Ncndl" preslika v značilke "Noun, Type:common, Gender:neuter, Number:dual, Case:locative". Specifikacije poleg formalnih delov vsebujejo tudi komentarje, bibliografijo itd.

Skupni del specifikacij podaja 14 definiranih kategorij MULTEXT, ki večinoma ustrezajo besednim vrstam, nekaj pa jih je uvedenih iz tehničnih razlogov. Vsaka kategorija ima namensko tabelo, ki določa njene atribute, njihove vrednosti in njihovo preslikavo v nize MSD. Za vsak par atribut-vrednost določi tudi jezike, za katere je ta par primeren.

Drugi glavni del specifikacij je sestavljen iz razdelkov, specifičnih za vsak posemezni jezik. Ti poleg uvoda vsebujejo tudi razdelke za vsako kategorijo s svojimi tabelami definicij vrednosti atributov. Te tabele so podobne skupnim tabelam v tem, da tudi podajo atribute in njihove vrednosti, vendar le tiste, ki so primerne za obravnavani jezik. Vendar pa te jezikovno specifične tabele lahko tudi redefinirajo položaj atributov v nizu MSD, kar vodi do veliko krajših in bolj berljivih oznak MSD za jezik.

Jezikovno specifične tabele lahko vsebujejo tudi informacije o lokalizaciji. To omogoča izražanje značilk in MSD-jev bodisi v angleščini ali v jeziku, ki je opisan, zaradi česar so bolj primerni za uporabo maternih govorcev jezika. Nenazadnje razdelek za določen jezik tudi našteje vse veljavne MSD-je, s čimer določi nabor oznak MSD za ta jezik. Množica dovoljenih MSD-jev je pomemben podatek, saj je z MSD-ji označen korpus mogoče samodejno preveriti glede na ta seznam, nabor oznak pa je mogoče tudi preoblikovati v različne druge formate.

Specifikacije v XML oz. TEI so opremljene s pripadajočimi programi XSLT, ki sprejmejo specifikacije kot vhodne podatke, običajno skupaj z določenimi parametri, in ustvarijo XML, HTML ali besedilni izhod, odvisno od slogovne datoteke. Na voljo so trije razredi transformacij. Prvi pomaga pri dodajanju novega jezika samim specifikacijam, drugi preoblikuje specifikacije v HTML za branje, tretji pa preoblikuje (in potrdi) seznam MSD-jev. Izhodi drugega in tretjega razreda transformacij so vključeni v distribucijo MULTEXT-East.

Povezave na specifikacije

- Smernice MULTEXT-East V6 na GitHub: https://github.com/clarinsi/mte-msd

- Smernice MULTEXT-East V6 v TEI: https://nl.ijs.si/ME/V6/msd/xml/
- Smernice MULTEXT-East V6 za branje: https://nl.ijs.si/ME/V6/msd/html/index.html
- Smernice MULTEXT-East V6 za slovenski jezik:
  - v TEI: https://nl.ijs.si/ME/V6/msd/xml/msd-sl.spc.xml
  - za branje: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html
  - oznake v tabeli TSV: https://nl.ijs.si/ME/V6/msd/tables/msd-human-sl.tbl
  - oznake kodirane kot strukture lastnosti v TEI: https://nl.ijs.si/ME/V6/msd/tables/msd-fslib2-sl.xml

Označevalne smernice

V tem razdelku so zbrane označevalne smernice za oblikoskladnjo MULTEXT-East. Smernice so razvrščene od nastarejše različice do zadnje, ažurne različice.

Različica 2.0 (25-02-2023)

projekt Razvoj slovenščine v digitalnem okolju
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]

Različica 1.0 za nestandardno slovenščino (21-12-2016)

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]

Različica 1.0 (2008)

projekt Sporazumevanje v slovenskem jeziku
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]

Smernice MULTEXT-East:

Smernice MULTEXT-East V6 na GitHub: https://github.com/clarinsi/mte-msd

Smernice MULTEXT-East V6 v TEI: https://nl.ijs.si/ME/V6/msd/xml/

Smernice MULTEXT-East V6 za branje: https://nl.ijs.si/ME/V6/msd/html/index.html
Smernice MULTEXT-East V6 za slovenski jezik:

- v TEI: https://nl.ijs.si/ME/V6/msd/xml/msd-sl.spc.xml
- za branje: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html
- oznake v tabeli TSV: https://nl.ijs.si/ME/V6/msd/tables/msd-human-sl.tbl
- oznake kodirane kot strukture lastnosti v TEI: https://nl.ijs.si/ME/V6/msd/tables/msd-fslib2-sl.xml

Reference in povezave

V tem razdelku so zbrane reference in povezave na projekte, v katerih se je označevalni sistem razvijal in uporabljal.

Projekti, na katerih se je označevalni sistem razvijal oz. uporabljal

MULTEXT-East - Multilingual corpora and text tools for Central and East European langauges

JOS - Jezikoslovno označevanje slovenskega jezika: metode in viri

Sporazumevanje v slovenskem jeziku

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

Razvoj slovenščine v digitalnem okolju

Učni korpusi z ročno pregledanimi oznakami MULTEXT-East

Arhar Holdt, Špela; et al., 2022, Training corpus SUK 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1747.

Lenardič, Jakob; et al., 2022, CMC training corpus Janes-Tag 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1732.

Krek, Simon; et al., 2019, Training corpus ssj500k 2.2, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1210.

Reference

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]

Erjavec, Tomaž. 2017. MULTEXT-East. V (Nancy Ide, James Pustejovsky, ur.): Handbook of linguistic annotation. pp. 441-462. Springer. DOI: 10.1007/978-94-024-0881-2_17.

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]

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]