

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 (attribute 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 "NcndI" 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 attribute, 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 attribute 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>
-

Revision #12

Created 3 March 2023 10:56:18 by Tina Munda

Updated 9 November 2023 15:15:45 by Tina Munda