

LIETUVIŲ KALBOS MORFOLOGIJA ATVIROJO KODO HUNSPELL PLATFORMOJE

Santrauka

Straipsnyje pateikiami naujos kartos kompiuterinės lietuvių kalbos morfologijos sukūrimo rezultatai. Šiam darbui atlikti buvo pasirinkta *Hunspell* platforma, kuri geriausiai atitiko iškeltus reikalavimus: 1) naudoti ir kurti tik atvirąjį kodą; 2) lietuvių kalbos specifines savybes gali koduoti tik duomenų reikšmės, bet ne tų duomenų forma ir interpretavimo būdas; 3) visas programinis kodas turi būti universalus, tikti bet kuriai kitai kalbai; 4) maksimaliai pasinaudoti kitoms pasaulio kalboms sėkmingai pritaikytais sprendimais. Be to, šios platformos pagrindu nesunkiai galima tikrinti rašybą daugelyje (per 50) taikomųjų programų atlikti žodžių morfologinę analizę bei sintezę, vykdyti intelektinę tekstinės informacijos paiešką ir pan.

Kalbos morfologiją *Hunspell* platformoje aprašo du tekstiniai failai. Viena iš jų (plėtinio **.aff**) yra užrašomos kaitybos taisyklės, o kita (plėtinio **.dic**) – žodžių formos su pradine morfologine informacija ir nuorodomis į vieną ar kelias kaitybos taisykles, kurios papildomai gali būti taikomos. Abiejų šių failų sudarymo pagrindas – *Dabartinės lietuvių kalbos gramatika* ir Vilniaus universiteto, Vytauto Didžiojo universiteto ir Lietuvos Respublikos Seimo tekstynai (iš viso apie 1,3 mlrd. žodžių). Buvo užrašyta maždaug 5 000 adresuojamų kaitybos taisyklių grupių ir atrinkta per 170 000 lemų. Vykiant lemų atranką buvo apsiribojama tiksliai taisyklinga šiuolaikine lietuvių kalba ir vengiama klaidingų, nelietuviškų, neteiktinų, įžeidžiančių, nebevarojamų žodžių. Neaiškiais atvejais buvo tariamasi su Valstybinės lietuvių kalbos komisijos specialistais.

Vidutinio šiuo metu internete atsirandančio teksto atpažįstamumas yra apie 98 proc., t. y. vidutiniškai 98 iš 100 žodžių morfologinis analizatorius interpretuoja teisingai.

Šis naujai sukurtas morfologinės analizės būdas buvo sėkmingai pritaikytas Vytauto Didžiojo universiteto Sintaksinės-semantinės analizės sistemoje bei Lietuvos Respublikos Seimo Teisės aktų registre. Ypač sėkmingas taikymas buvo *Solr/Lucene* dokumentų indeksavimo ir paieškos sistemoje, kur pavyko ne tik tiksliai indeksuoti ir surasti visas norimas žodžių formas, bet ir pagreitinti indeksavimą šimtus kartų. Tokius rezultatus pavyko pasiekti pakeitus morfologijos **.aff**, **.dic** aprašą baigtiniu būsenų keitikliu (FST – *Finite State Transducer*).

Toliau plėtojant šiuos darbus reikėtų parengti bent dvi žodyno versijas: norminę, tinkamą rašybai tikrinti, ir pilną (norminė ir nenorminė kalba), tinkamą taikomiesiems gyvosios kalbos analizės uždaviniams spręsti.

VIRGINIJUS DADURKEVIČIUS

Vilniaus universiteto Taikomųjų mokslų institutas

M. K. Čiurlionio g. 29, 03100 Vilnius

virginijus.dadurkevicius@tmi.vu.lt

dadurka@gmail.com