Ekstraktmorfoloogia meetodiga tuletatud keeletehnoloogia vadja sõnavara näitel

Kristian Kankainen

2019

Sisukord

1 Sissejuhatus

Magistritöö loob viisi ehitada arvutimorfoloogia puhtalt lekseemide sõnavormide esitamise teel ning teisendada ehitatud arvutimorfoloogilise mudeli automaatselt kahte keeletehnoloogilisse raamistikku.

Magistritöö kasutab loodud süsteemi selleks, et kirjeldada vadja keele normatiivsed morfoloogilised tüüpsõnad.

Tööd ajendab mõtteviis minimeerida tööd: loodud normatiivne morfoloogiline tüübistik on aluseks automaatselt tuletatud keeletehnoloogiale, kui normatiiv muutub, muutub ka keeletehnoloogia. Töö paneb leksikaalse ressursi esikohale ja kõik leitud sisulised vead õiendatakse otse ressursis, mitte keeletehnoloogilistes tarkvarades eraldi.

1.1 Teoreetilised lähtekohad

Morfeemi ei käsitleta siin töös levinust lingvistilisest seisukohast kui *väikseimat tähenduslikku üksust*, vaid klassikalistele paradigmaatilistele lähenemistele omaselt kui *mistahes tähtkoostise muutust, millega kaasneb tähenduslik muutus* (Beard, 1987, Beard, 1995).

2 Ekstraktmorfoloogia meetod

See osa kirjeldab töös rakendatud meetodit. Töö kasutab ekstraktmorfoloogiat kaheks otstarbeks, esiteks vadja keele morfoloogiliste tüüpsõnade väljaselgitamiseks ja kirjeldamiseks ja teisalt programmkoodi automaatseks tuletamiseks saadud kirjelduse põhjal.

Meetodi kaks rakendust on lähemalt kirjeldatud järgnevates peatükkides Vadja morfoloogiliste tüüpsõnade analüüs ja Programmkoodi tuletamine vastavalt.

2.1 Sissejuhatus

3 Vadja morfoloogiliste tüüpsõnade analüüs

See osa kirjeldab ekstraktmorfoloogiaga leitud vadja keele morfoloogilisi tüüpsõnu ja analüüsib nende vastavust vadja keele grammatikatega ja ajaloolise morfoloogiaga.

4 Programmkoodi tuletamine

Programmkoodi tuletamise all peetakse siin töös silmas mistahes protsessi, mille käigus tuletatakse mingi üldisema kirjelduse põhjal programmkoodi ühe või mitme konkreetse programmeerimiskeskkona jaoks.

Üldine kirjeldus (või teisisõnu ontoloogia) kirjeldab faktuaalselt *mida* ning tuletatud programmkood kirjeldab konkreetselt *kuidas* seda teadmist rakendada.

Töös kasutatakse keskseks kirjelduseks leksikaalset ressurssi, mille peamine osa koosneb ekstraktmorfoloogiaga leitud tüüpsõnade mallidest.

Keskse kirjelduse leksikaalset ressurssi hoitakse rahvusvahelise standardi vormingus *Lexical Markup Framework* (ISO/TC 37/SC 4, 2007).

Programmkoodi tuletavad nn generaatorid. Töös esitatakse kaht generaatorit, üks programmeerimiskeele Grammatical Framework jaoks ning teine keeletehnoloogilise taristu Giellatekno integreerimise jaoks.

4.1 Keskse kirjelduse hoidmine Lexical Markup Framework vormingus

Sissejuhatav tekst, mis on e-sõnastike rahvusvaheline standard Lexical Markup Framework (ISO/TC 37/SC 4, 2007) ja milleks seda kasutatakse.

LMF koosneb mitmest laiendimoodulist (vt nt Francopoulo, 2013), millest siinne töö kasutab kahte: morfoloogia moodul (*LMF Morphology Extension*) ja morfoloogiliste paradigmade moodul (*LMF Morphological Pattern Extension*).

4.2 Grammatical Framework morfoloogiakomponent

Mis on programmeerimiskeel Grammatical Framework ja milleks seda kasutatakse.

4.3 Giellatekno taristuga integreerimine

Mis on keeletehnoloogiline taristu Giellatekno ja milleks seda kasutatakse. Kes seda kasutavad.

5 Kokkuvõte

Magistritöö on kirjeldanud süsteemi, millega on ühelt poolt defineeritud vadja keele normatiivne morfoloogia ja mille põhjal teisalt tuletatakse automaatselt morfoloogiline keeletehnoloogia.

Morfoloogiline normatiiv põhineb Heinike Heinsoo läbiviidud keelekümbluskoolis Ämmesse Vunukassaa õpetatud keelel.

Saadud morfoloogilist tüübistikku on analüüsitud vadja keele grammatikatega ja põhjendatud ajaloolise morfoloogiaga.

6 Põhimõisted ja lühendid

Siin loetletakse töös kasutatud mõisted ja lühendid koos nende tähendustega.

7 Kirjandus

Beard, Robert (1987). „Morpheme order in a lexeme/morpheme-based morphology“. *Lingua* 72.1, lk. 1–44.

— (1995). *Lexeme-morpheme Base Morphology: A General Theory of Inflection and Word Formation*. SUNY Series in Linguistics. OCLC: 940540414. State University of New York Press.

Francopoulo, Gil (2013). *LMF lexical markup framework*. London; Hoboken, NJ: ISTE Ltd ; John Wiley & Sons. ISBN: 1-84821-430-8 978-1-84821-430-9.

ISO/TC 37/SC 4 (30. juuni 2007). *Language resource management Lexical markup framework (LMF)*. 24613:2007 Rev.14. ISO. URL: `http://lirics.loria.fr/doc_pub/LMF_revision_14.pdf` (vaadatud 13.06.2017).

8 The use of Extract Morphology for Automatic Derivation of Language Technology for Votic

An English language summary of this work.