Ekstraktmorfoloogia meetodiga tuletatud keeletehnoloogia vadja sõnavara näitel

Kristian Kankainen

2019

Sisukord

1 Sissejuhatus

Magistritöö esimene eesmärk on luua H. Heinsoo Sõnakopittöjas esitatud sõnavarast morfoloogiline sõnastik, mis sisaldab sõnavara kõiki muutvorme. Selleks vajalik arvutimorfoloogiline kirjeldus ehitatakse sellisel moel, et see taandub tüüpsõnade muutvormitabelite esitamisele, mitte grammatiliste reeglite esitamisele. Niiviisi ehitatud teooriavaba(m) arvutimorfoloogiline kirjeldus võimaldab luua erinevaid keeletehnoloogiaid automaatselt programmkoodi tuletamise teel. Esitatakse kolme tehnoloogia automaatset tuletamist: 1) ühe keeletehnoloogilise taristusse integreerimise kaudu õigekirjakontrollija, 2) vadja keele arvutimorfoloogia moodul ühe loomulike keelte grammatikate koostamiseks mõeldud programmeerimiskeelele ja 3) morfoloogia tehnoloogiaülene kirjeldus ühe rahvusvahelise standardi abil.

Kuna kõik tuletatud keeletehnoloogia edaspidine täiendamine ja täpsustamine käib ainult lekseemide muutvormitabelite täiendamise ja täpsustamise kaudu, peab esimese eesmärgi juurde lisama seda, et magistritöös loodud leksikograafiline süsteem võimaldab keeleaktivistide rühmal töötada oma sõnavara ja keeletehnoloogia kallal edaspidi ka ilma spetsialistist keeleteadlase ja keeletehnoloogi abil. Kas seda vadja keele puhul ka juhtub, jääb tuleviku näidata.

Magistritöö teine eesmärk on analüüsida leitud tüüpsõnad mitmel viisil: 1) kirjeldada nende morfofonoloogiat keeleajalooliste arengute taustal, 2) leida tüüpsõnade põhivormid ja analoogiavormid, 3) esitada üks võimalik muuttüüpide süsteem ja võrrelda seda seni esitatutega ja viimalt 4) analüüsida muuttüüpide produktiivsust.

—

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 muu-

tub, muutub ka keeletehnoloogia. Töö paneb leksikaalse ressursi esikohale ja kõik leitud sisulised vead õiendatakse otse ressursis, mitte keeletehnoloogilistes tarkvarades eraldi.

2 Teoreetilised lähtekohad

Kuna töö opereerib arvutilingvistika, deskriptiivse ja dokumentaalse lingvistika ääremail, peame selgitama töö teoreetilised lähtekohad. Siinsele kompendiumiks on ka põhimõisted seletatud pt 7 Põhimõisted ja lühendid.

2.1 Vadja kirjakeel ja normatiiv

Vadja keelele ei loodud kirjandust 1930-ndateil, nagu seda tehti Nõukogude Liidus näiteks karjala, vepsa ja isuri keele jaoks.

Siiski on vadja keelel hulganisti lingvistilisi kirjeldusi, nagu grammatikaid (mh Ahlqvist 1856; Airila 1934; Tsvetkov 2008; Ariste 1968; Маркус ja Рожанский 2011), sõnaraamatuid (mh Tsvetkov 1995; Ariste 1943; Laakso 1989; Raag 1982; Pomberg ja H. Heinsoo 1991; Grünberg *et al.* 2013; Heinike Heinsoo 2015) ja ka etnograafilisi töid (mh Kass 1961; Mälk 1977).

Kirjeldused ei aita siiski kaasa tänapäeva keeleõppija küsimustele *kuidas kirjutatakse sõna* TÜTTÖ *mitmuse omastavas?*. Selleks on vaja tänapäevase vadja keele morfoloogia standardiseerimist ehk normatiivset kirjeldust.

Käesolev töö ei pürgi looma lõplikku normatiivi, kuivõrd ta loob süsteemi, mis oskab vastata morfoloogilistele küsimustele. Aga loodud süsteemi peamine eesmärk on siiski võimaldada muuta ja jätkata tööd normatiivi arendamiseks ja mille ümber saaks keeleaktivistid ise koonduda, ilma et selleks oleks niivõrd vaja ei lingvistilist ega keeletehnoloogilist spetsialisti.

Püüd luua vadja morfoloogiale normatiivne alus lihtsustab paljudele küsimustele vastusi leida, nt mis käändeid arvestada. Siiski on tööga loodud *keeletehnoloogia tuletamise süsteem* avatud ka teistsugustele lähenemistele keeleainesele.

2.2 Ortograafia

On järgitud Heinsoo loodud ortograafiat mille jaoks on Kankainen teinud vadja klaviatuuripaigutise (Kankainen, ilmumas). Vadja erinevatest kirjaviisidest on kirjuta-

nud Ernits 2010 ja erinevatest kirjakeele loomise pürgimistest Ernits 2006.

2.3 Morfofonoloogia

Tavaliselt jagatakse arvutilingvistikas morfoloogia ja fonoloogia eraldi nii, et morfoloogia tasandil on abstraktne esitus, nn morfofoneemid, mille pindesitused tulenevad eraldi fonoloogilistest reeglitest.

Niiviisi saaks esitada mõlemad fonoloogilised vormid *tšiuttoa* ja *tüttöä* ühe ja sama morfoloogilise kujuga TŠIUTTO+A ja TÜTTÖ+A. Kusjuures käändelõpu +A pindesitus *a*-na või *ä*-na sõltuks vastavalt sellest, kas lemmas esineb tagapoolsed või eespoolsed vokaalid.

See töö ei arvesta morfofonoloogilise tasandiga. Peatükis 4.3 näidatakse üht võimalikku viisi koondada tüüpsõnu kokku abstraktsemal tasandil, mis mingil määral arvestab ka morfofonoloogilisi reeglipärasusi.

2.4 Klassikaline paradigmaatiline morfoloogia

Sõna kui selle vormide tervik; pedagoogiline praktika ja paradigma üldistuse ülekantavus uutele sõnadele Matthews 1991. Matthewski jätab mudeli vormipõhiseks ja mitte morfeemipõhiseks, selle kohta edasi järgmises allosas.

2.5 Morfeemi staatus ja definitsioon

Morfeemi ei käsitleta siin töös levinud lingvistilisest seisukohast kui *väikseimat tähenduslikku üksust*, vaid klassikalistele paradigmaatilistele lähenemistele omaselt kui *mistahes tähtkoostise muutust, millega kaasneb tähenduslik muutus* (Beard 1987; Beard 1995). Morfeemipõhist suunda ajab nt Stump 2001.

2.6 Muuttüüp, tüüpsõna ja muutkond

Eesti traditsiooni järgi on muuttüüp tüüpsõnast üldisem. Kuidas siin töös terminoloogiliselt ümber käia, kas *muuttüüp* või *tüüpsõnamall*?

Muuttüübistik sõltub selle aluseks võetud klassifikatsioonist, ekstraktmorfoloogiat võiks vaadata kui lihtsalt üht väga formaalselt defineeritud muuttüübistikku.

Huldenil on omakorda üks väga formaalne viis, kuidas vähendada ekstraktmorfoloogiaga leitud muuttüüpide arvu. Kas see on hoopis muuttüübistik?

3 Ekstraktmorfoloogia meetod

See osa kirjeldab töös rakendatud ekstraktmorfoloogia meetodit. Töö kasutab ekstraktmorfoloogiat kaheks otstarbeks, esiteks vadja keele morfoloogiliste tüüpsõnade väljaselgitamiseks ja kirjeldamiseks ja teisalt programmkoodi automaatseks tuletamiseks saadud kirjelduse põhjal. Neid kahte rakendust kirjeldatatakse lähemalt vastavates peatükkides *Vadja morfoloogiliste tüüpsõnade analüüs* ja *Programmkoodi tuletamine*.

3.1 Sissejuhatus

Ekstraktmorfoloogia on juhendatud masinõppe meetod, mis üldistab lekseemide muutvormitabeleid ja eraldab neist tüüpsõnamallid. See on *juhendatud*, sest sisendiks olevad muutvormitabelid peavad olema korrektselt koostatud.

Selles töös käsitletakse meetodi abil saadud mudelit siiski pigem lihtsa kirjeldusena. See on tüüpsõnakirjeldus, mis on osa sõnastikust – lekseemi paradigma kirjeldusena. Ja sellest kirjeldusest

Tüüpsõnamall koosneb muutvormide mallidest ja vastab seega morfoloogilise paradigma mõistele. Tüüpsõnamalli abil on võimalik moodustada ka tundmatu sõna kõik muutvormid. Kuna kaks või enam lekseemi võivad jagada üht ja sama tüüpsõnamalli (s.o kuuluda sama paradigmasse), on võimalik ekstraktmorfoloogia meetodiga üldistada lekseemide iseärasusi ja luua nendest tüüpsõnade produktiivsuse mudeli. Produktiivsusmudeliga on võimalik ennustada uue ja tundmatu sõnavormi kuuluvust ühe või teise tüüpsõna alla.

Veel ilma detailidesse takerdumata näitlikustatakse siinkohal lugejale meetodi sisendit ja väljundit. Sisendiks on ühe lekseemi muutvormitabel tervikuna (vt tabel 1). Väljundiks on meetodi poolt leitud tüüpsõnamall (vt tabel 2). Tabelitele viidatakse alljärgnevas tekstis mitmel korral.

muutvorm	tunnused
katto	SG NOM
katod	PL NOM
kato	SG GEN
kattoi	PL GEN
kattoje	PL GEN
kattoa	SG PART
kattoi	PL PART
kattoite	PL PART
kattose	SG ILL
kattoise	PL ILL
kattoz	SG INE
kattoiz	PL INE
katosse	SG ELA
kattoisse	PL ELA
katolle	SG ALL
kattoille	PL ALL
katol	SG ADE
kattoil	PL ADE
katolte	SG ABL
kattoilte	PL ABL
katossi	SG TRAN
kattoissi	PL TRAN
kattossaa	SG TERM
kattoissaa	PL TERM
katoka	SG COM
kattoika	PL COM

Tabel 1: Sisendi muutvormide tabel koos morfoloogiliste tunnustega.

ühisosajada	muutvormimall	tunnused
kat t o	x_1 + t + x_2	SG NOM
kat o d	x_1 + x_2 + d	PL NOM
kat o	x_1 + x_2	SG GEN
kat t o i	x_1 + t + x_2 + i	PL GEN
kat t o je	x_1 + t + x_2 + je	PL GEN
kat t o a	x_1 + t + x_2 + a	SG PART
kat t o i	x_1 + t + x_2 + i	PL PART
kat t o ite	x_1 + t + x_2 + ite	PL PART
kat t o se	x_1 + t + x_2 + se	SG ILL
kat t o ise	x_1 + t + x_2 + ise	PL ILL
kat t o z	x_1 + t + x_2 + z	SG INE
kat t o iz	x_1 + t + x_2 + iz	PL INE
kat o sse	x_1 + x_2 + sse	SG ELA
kat t o isse	x_1 + t + x_2 + isse	PL ELA
kat o lle	x_1 + x_2 + lle	SG ALL
kat t o ille	x_1 + t + x_2 + ille	PL ALL
kat o l	x_1 + x_2 + l	SG ADE
kat t o il	x_1 + t + x_2 + il	PL ADE
kat o lte	x_1 + x_2 + lte	SG ABL
kat t o ilte	x_1 + t + x_2 + ilte	PL ABL
kat o ssi	x_1 + x_2 + ssi	SG TRAN
kat t o issi	x_1 + t + x_2 + issi	PL TRAN
kat t o ssaa	x_1 + t + x_2 + ssaa	SG TERM
kat t o issaa	x_1 + t + x_2 + issaa	PL TERM
kat o ka	x_1 + x_2 + ka	SG COM
kat t o ika	x_1 + t + x_2 + ika	PL COM

Tabel 2: Väljundi tüüpsõnamall (kusjuures $x_1 = kat$ ja $x_2 = o$ vastab mallist leitud ühisosajadale).

4 Vadja morfoloogiliste tüüpsõnade analüüs

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

4.1 Ekstraktmorfoloogiaga leitud tüüpsõnad

See alaosa loendab leitud tüüpsõnad sõnaliigiti. Analüüsitakse tüüpsõnade alla kuuluvaid sõnu struktuurselt (kui mitu silpi, silpide struktuur).

Analüüsidest on võimalik luua arvutikirjeldus hüpoteetilise vadjakeelse sõna üle õigekirjakontrollija jaoks.

4.2 Põhivormid ja analoogiavormid

Selles osas selgitatakse välja vadja keele tüüpsõnade põhi- ja analoogiavormid sõnaliigiti. Seda püütakse teha formaalselt põhinedes vaid ekstraktmorfoloogiaga leitud tüüpsõnamallidele.

M. Erelt, T. Erelt ja Ross 2007 järgi "[p]õhivormid on need vormid, mida pole võimalik teiste vormide alusel tuletada ning mille moodustamiseks tuleb iga sõnatüübi korral anda vastavad reeglid." ja "[a]naloogiavormid on vormid, mida saab moodustada mingi põhivormi analoogial."

4.2.1 Käändsõnad

4.2.2 Tegusõnad

4.3 Üks võimalik muuttüüpide süsteem

Silfverberg ja Hulden (2018) on kirjeldanud üht võimalikku formaalset viisi, kuidas ekstraktmorfoloogia tüüpsõnade arvu vähendada. Siin alaosas esitatakse selle põhjal loodud vadja muuttüübistik ja võrreldatakse seda Eesti muuttüüpide traditsiooniga.

Eesti muuttüüpide traditsioonist on kirjutanud mh Viks 2018.

4.4 Muuttüüpide produktiivsus

Kristiina Kross (Ross) nimetab produktiivsuseks "mingi morfoloogilise nähtuse võimet allutada endale uusi sõnu" (Kross 1984). Siin allosas seatakse eelmises osas leitud muuttüübid pingeritta selle järgi, kui mitu tüüpsõna nendele allub.

Kas selleks on vaja defineerida, mis on *uus sõna*? Näiteks kõik uuemad vene keele laenud.

Kas produktiivsuse pingerida on võimalik jagada mingi kriteeriumi järgi avatuteks ja suletuteks muuttüüpideks?

5 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 Giella keeletehnoloogilise taristu integreerimise jaoks. Generaatorid on kirjutatud programmeerimiskeeles XQuery.

5.1 Keskne kirjeldus Lexical Markup Framework vormingus

Sissejuhatav tekst, mis on e-sõnastike ja leksikaalsete andmebaaside rahvusvaheline standard Lexical Markup Framework (ISO/TC 37/SC 4 2007) ja milleks seda kasutatakse. (märksõnu: semantika eeldefineeritud märgenduskeel; koostöövõime)

Standardi märgenduskeel koosneb mitmest eriotstarbelisest laiendimoodulist (vt nt Francopoulo 2013). Siinne töö kasutab kahte: morfoloogia moodul (*LMF Morphology Extension*) ja morfoloogiliste paradigmade moodul (*LMF Morphological Pattern Extension*).

Morfoloogiamooduli eesmärgiks on kirjeldada morfoloogiat mahu kaudu, s.o kirjeldada lekseemi loendades kõiki selle muutvorme.

Morfoloogiliste paradigmade mooduli eesmärgiks on seevastu kirjeldada sisu kaudu, s.o kirjeldada neid kriteeriume ja reegleid, millega saab moodustada kõik ühe lekseemi muutvormid. Selles töös kirjeldatakse ekstraktmorfoloogia tüüpsõnamalle antud mooduliga.

Sama nähtuse kirjeldamine nii mahus kui ka sisus võib tunduda liigsena, ent niiviisi võimaldatakse rohkem informatsiooni hoidmist.

Näiteks võib iga lekseemi muutvormi kohta hoida informatsiooni nende reaalsetest korpusesinemustest. Niiviisi on võimalik klassifitseerida tüüpsõnade teoretiseeringutaset, kui ühe ja sama tüüpsõna alla kuuluvate lekseemide korpusleiud kinnitavad igat selle muutvormi, ei ole see teoretiseeritud.

Peale sõnaartiklite ja morfoloogilise informatsiooni hoitakse leksikaalses ressursis ka globaalset informatsiooni, nagu keele nimetus ja kood.

5.1.1 Sõnaartiklite esitamine LMFis

Iga sõnaartikkel ehk leksikaalne kirje kannab informatsiooni lekseemi sõnaliigi kohta, selle valitud lemma vorm ning morfoloogiamooduliga esitatud muutvormitabeli.

5.1.2 Tüüpsõnamallide esitamine LMFis

5.2 Grammatical Framework morfoloogiakomponent

Mis on see, mida mina teen. Seejärel, mis on programmeerimiskeel Grammatical Framework ja milleks seda kasutatakse.

Morfoloogiakomponendi programmkood on jaotatav kaheks tükiks, leksikoniks ja muuttüüpide funktsioonid. Järgnevalt neist detailsemalt. Viimases alaosas on arutelu GFide võimalustest ja edasiarendusvõimalustest.

5.2.1 Leksikon

5.2.2 Muuttüübid

5.2.3 Arutelu

Loodud morfoloogiakomponenti on kasutatud interaktiivses vadja-vene-vadja vestmikus.

5.3 Integreerimine Giella-taristuga

Keeletehnoloogilise taristuga Giella integreeritakse selles töös peamiselt selleks, et saada kätte õigekirjakontrollija. Giella-taristu koosneb veel võimalustest. Taristut kasutavad peamiselt Giellatekno ja Divvun.

Integreerimine on jagatav kaheks peamiseks osaks: leksikoni integreerimeerimine ja tüüpsõnamallide integreerimine. Seejärel kirjeldatakse taristu poolt loodud õigekirjakontrollija tööpõhimõtet ja lõpetuseks on arutelu.

5.3.1 Leksikon

"Formally, the lexc language is a kind of right-recursive phrase-structure grammar." ja "A lexc description compiles into a standard Xerox finite-state network, either a simple automaton or a transducer." (Beesley ja Karttunen 2003, lk 203).

Kuigi lexc fraasistruktuurigrammatikatega on võimalik paradigmasid (tüüpsõnamalle) mudeldada, ja tavaliselt selleks seda kasutataksegi Giella taristus, võtab see töö teise lähenemisnurga ja lihtsustab võimalikult palju leksikoni struktuuri.

Leksikon koosneb selles töös ainult kahest andest: *lemma* ja *tüüpsõna*.

5.3.2 Muuttüübid

Paradigmade ehk tüüpsõnamallide esitus FST formalismis põhineb suuresti Forsbergi ja Huldeni (2016) tööle.

Paradigmad esitatakse relatsioonidena sõnavormi ja lemma koos analüüsiga vahel. Sellised relatsioonid sisaldavad lõpmatut hulka sõnalemmasid, millest mõistagi pole suurem osa vadjakeelsed. Mis on siiski tähtis, on see, et relatsioonid mudeldavad paradigmasid.

Sõnade lõpmatu hulk piiratakse leksikonis antuga ja niiviisi saadakse leksikonis sisalduvate sõnade kõik sõnavormid. Nendest ja ainult nendest sõnavormidest koosnebki esialgne vadja õigekirjakontrollija.

5.3.3 Õigekirjakontrollija

Eelnevalt kirjeldatud integreerimine Giella-taristusse võimaldab taristul luua õigekirjakontrollija. Mis on õigekirjakontrollija, kus seda kasutatakse ja mida see kontrollib?

5.3.4 Arutelu

Loodud õigekirjakontrollija on eesmärgipäraselt jäetud lihtsakoeliseks. See märgib kõik sõnad valeks, mis ei sisaldu sõnastikus. See on lühiajaliseks kasutamiseks ja mõeldud ärgitama kasutajaid ise pakkuma täiendusi ja sõnaloomet vadja sõnastikusse.

6 Kokkuvõte

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

Morfoloogilise normatiivi vajadust ajendab Heinike Heinsoo läbiviidud kursused keelekümbluskoolis Ämmesse Vunukassaa ja normatiiv on hõlpsasti muudetav-parendatav ilma programmeerimisoskusteta.

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

—

Töö keskseks osaks on ekstraktmorfoloogiameetodiga saadud tüüpsõnakirjeldused. Kirjeldused kodeeritakse koos sõnastikuga ümber standardsesse vormingusse ja saadud leksikaalse ressursi järgi tuletatakse automaatselt programmkoodi kahe keeletehnoloogilise platvormi jaoks, ja tagatakse seega vadja keele tugi nendes platvormides.

Niivisii kasutatakse ekstraktmorfoloogia meetodit kasutajaliidesena, mille abil koostatakse arvutimorfoloogia ainult tüüpsõnade muutvormitabeleid sedastades – mitte programmeerides.

Magistritöös esitatud töövoog paneb leksikaalse ressursi kesksele kohale ja tuletatud tehnoloogia sellest teiseseks. Uue sõnavara ja vigade parandused tehakse ressursis, mitte mitmes tehnoloogias eraldi.

Kuna nii tüüpsõnade kirjeldused, kui ka ülejäänud sõnastik kodeeritakse rahvusvahelise standardi Lexical Markup Framework vormingusse, tagatakse võimaluse ressursi pikaajaliseks arhiveerimiseks. Leksikaalne ressurss on loetav ja arusaadav palju kauem, kui seda on programmeerimiskood.

Viimase tõttu püüab magistritöö ühendada arvutuslingvistika ja dokumenteeriva lingvistika valdkondi.

7 Põhimõisted ja lühendid

Siin loetletakse töös kasutatud mõisted ja lühendid koos nende tähendustega.

Konkatenatsioon ehk $\oplus$ on tähtede ja tähtjärjendite lükkimine teine-teise järele, et moodustada uus tähtjärjend. Näiteks *aa* $\oplus$ *be* moodustab *aabe*. 16

Lemma on suvaliselt valitud grammatiliste tunnuste komplekt, mida kasutatakse lekseemi viitamiseks. 16

Mikrostruktuur on sõnastiku sõnaartikli sisemine struktuur. 16

Muutvormimall kirjeldab üksiku muutvormi koostamisskeemi ja kannab selle grammatilised tunnused. On integraalne osa tüüpsõnamallist. Koostamisskeem koosneb muutujatest ja konstantidest, mille tähtkoostised lükitakse üks-teise järele. Muutujate tähtkoostised võivad olla mingil moel piiratud. 16

Tehniline tüvi on tähtkoostiste järjend, millega saab tüüpsõnamalli muutvormide muutujad asendada elik väärtustada ja niiviisi koostada ühe konkreetse sõna kõik vormid. 16

Tüüpsõnamall on ekstraktmorfoloogiaga leitud tüüpsõna paradigma kirjeldus, mis koosneb iga muutvormi koostamismallidest ehk muutvormimallidest. Tüüpsõnamall on relatsioon tehnilise tüve ja kõigi selle paradigmasse kuuluvate muutvormide vahel. 7, 16

8 Kirjandus

Ahlqvist, August (1856). *Wotisk grammatik jemte språkprof och ordförteckning: (Föredr. d. 15 Oktober 1855)*. [Helsingfors: s.n. 162 lk. kokku.

Airila, Martti (1934). *Vatjan kielen taivutusoppi. 1: Nominien taivutus*. Vähäisiä kirjelmiä 87. Helsinki: Suomalaisen kirjallisuuden seura. 55 lk. kokku.

Ariste, Paul (1943). *Vadja lemmüz: mõningaid vadja sõnaseletusi: avec résumés français*. Helsinki: s.n. 1 lk. kokku.

— (1968). *A grammar of the Votic language*. Indiana University publications. Uralic and Altaic series vol. 68. Bloomington : The Hague: Indiana University ; Mouton. 121 lk. kokku.

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.

Beesley, Kenneth R ja Lauri Karttunen (2003). *Finite state morphology*. Stanford, Calif.: CSLI Publications. ISBN: 1-57586-433-9 978-1-57586-433-4 1-57586-434-7 978-1-57586-434-1.

Erelt, Mati, Tiiu Erelt ja Kristiina Ross (2007). *Eesti Keele Käsiraamat*. 3., täiend. tr. Tallinn: Eesti Keele Sihtasutus. 726 lk. kokku. ISBN: 978-9985-79-210-0.

Ernits, Enn (2006). „Vadja liikumisest ja kirjakeelest“. *Keel ja Kirjandus* 49.1, lk. 85–87. URL: https://www.digar.ee/viewer/et/nlib-digar:81648/143905/page/85.

— (2010). „Vadja kirjaviisist ja sõnaloomest“, lk. 17.

Forsberg, Markus ja Mans Hulden (2016). „Learning Transducer Models for Morphological Analysis from Example Inflections“. *Proceedings of StatFSM*. Association for Computational Linguistics, lk. 42. URL: http://anthology.aclweb.org/W16-2405.

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.

Grünberg, Silja *et al.*, toim. (2013). *Vadja keele sõnaraamat =: Vaďďaa tšeelee sõna-tširja = Словарь водского языка*. 2., täiend. ja parand. tr. Tallinn: Eesti Keele Sihtasutus. 1823 lk. kokku. ISBN: 978-9985-79-553-8.

Heinsoo, Heinike (2015). *Vad'd'a sõnakopittõja*. Koostöös Helena Miettinen *et al.* Helsinki ; Tartu: Mooses Putron muistosäätiö, Tallinna Raamatutrükikoda). 182 lk. kokku. ISBN: 978-952-93-5025-4.

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).

Kass, Asta (1961). „Käsitöö- ja rõivastusalane sõnavara vadja keeles“. Tartu.

Kross, Kristiina (1984). *Eesti noomeni muutmistüüpide produktiivsus: soome-eesti kontrastiivseminar, (Helsingi, 12.-20. okt. 1984): [ettekanne]*. Koostöös Keele ja Kirjanduse Instituut ja Eesti NSV Teaduste Akadeemia. Ars grammatica KKI-26. Tallinn: Eesti NSV Teaduste Akadeemia. 40 lk. kokku.

Laakso, Johanna, toim. (1989). *Vatjan käänteissanasto*. Lexica Societatis Fenno-Ugricae 22. Helsinki: Suomalais-ugrilainen seura. 103 lk. kokku. ISBN: 978-951-9403-21-2.

Matthews, Peter Hugoe (1991). *Morphology*. 2nd ed. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press. 251 lk. kokku. ISBN: 978-0-521-41043-4 978-0-521-42256-7.

Mälk, Vaina (1977). *Vadja vanasõnad eesti, soome, karjala ja vene vastetega*. Koostöös Keele ja Kirjanduse Instituut. Tallinn: Eesti Raamat. 404 lk. kokku.

Pomberg, Merle ja H. Heinsoo (1991). „Vadja tööriistade ja tarbeesemete nimetused Eesti Rahva Muuseumis: kursusetöö“. Tartu.

Raag, Virve (1982). *A dictionary of Votic*. Uppsala: Fenno-ugrica suecana. 230 lk. kokku.

Silfverberg, Miikka, Ling Liu ja Mans Hulden (20. august 2018). „A Computational Model for the Linguistic Notion of Morphological Paradigm“, lk. 12.

Stump, Gregory T (2001). *Inflectional morphology a theory of paradigm structure*. Cambridge; New York: Cambridge University Press. ISBN: 978-0-511-01378-2 978-0-521-78047-6 978-0-511-48633-3. URL: http://dx.doi.org/10.1017/CBO9780511486333 (vaadatud 19.07.2015).

Tsvetkov, Dmitri (1995). *Vatjan kielen Joenperän murteen sanasto*. Helsinki: Suomalais-ugrilainen seura ;Kotimaisten kielten tutkimuskeskus. ISBN: 978-951-9403-83-0.

— (2008). *Vadja Keele Grammatika*. Koostöös Jüri Viikberg, Ada Ambus ja Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus. 169 lk. kokku. ISBN: 978-9985-79-216-2.

Viks, Ülle (2018). *Muuttüübid eesti ükskeelsetes sõnastikes*. URL: https://www.eki.ee/teemad/tyybijutt.html (vaadatud 04.03.2018).

Маркус, Елена Борисовна ja Федор Иванович Рожанский (2011). *Современный водский язык: тексты и грамматический очерк. Том 2, Грамматический очерк и библиография: [в 2-х томах]*. Koostöös Институт языкознания (Moskva). Санкт Петербург: Нестор-История. 381 lk. kokku. ISBN: 978-5-98187-834-3.

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

An English language summary of this work.

```
<LexicalEntry morphologicalPatterns="asKatto">
  <feat att="partOfSpeech" val="nn"/>
  <Lemma>
    <feat att="writtenForm" val="katto"/>
  </Lemma>
  <WordForm>
    <feat att="writtenForm" val="katto"/>
    <feat att="grammaticalNumber" val="singular"/>
    <feat att="grammaticalCase" val="nominative"/>
  </WordForm>
  <WordForm>
    <feat att="writtenForm" val="katod"/>
    <feat att="grammaticalNumber" val="plural"/>
    <feat att="grammaticalCase" val="nominative"/>
  </WordForm>
</LexicalEntry>
```

Joonis 1: Sõnaartikli *katto* esitamine LMFis (muutvormid kajastatud vaid osaliselt).

```
<MorphologicalPattern>
  <feat att="id" val="asTšiutto"/>
  <feat att="partOfSpeech" val="nn"/>
  <TransformSet>
    <GrammaticalFeatures>
      <feat att="grammaticalNumber" val="singular"/>
      <feat att="grammaticalCase" val="nominative"/>
    </GrammaticalFeatures>
    <Process>
      <feat att="operator" val="addAfter"/>
      <feat att="processType" val="pextractAddVariable"/>
      <feat att="variableNum" val="1"/>
    </Process>
    <Process>
      <feat att="operator" val="addAfter"/>
      <feat att="processType" val="pextractAddConstant"/>
      <feat att="stringValue" val="t"/>
    </Process>
    <Process>
      <feat att="operator" val="addAfter"/>
      <feat att="processType" val="pextractAddVariable"/>
      <feat att="variableNum" val="2"/>
    </Process>
  </TransformSet>
  <TransformSet>
    <GrammaticalFeatures>
      <feat att="grammaticalNumber" val="plural"/>
      <feat att="grammaticalCase" val="nominative"/>
    </GrammaticalFeatures>
    <Process>
      <feat att="operator" val="addAfter"/>
      <feat att="processType" val="pextractAddVariable"/>
      <feat att="variableNum" val="1"/>
    </Process>
    <Process>
      <feat att="operator" val="addAfter"/>
      <feat att="processType" val="pextractAddVariable"/>
      <feat att="variableNum" val="2"/>
    </Process>
    <Process>
      <feat att="operator" val="addAfter"/>
      <feat att="processType" val="pextractAddConstant"/>
      <feat att="stringValue" val="d"/>
    </Process>
  </TransformSet>
<MorphologicalPattern>
```

Joonis 2: Tüüpsõnamalli `tšiutto` (mille alla kuuluvad mh *tšiutto* ja *katto*) esitus LMFis. Esitus mudeldab muutvormimalle $x_1 \oplus \mathbf{t} \oplus x_2$ ning $x_1 \oplus x_2 \oplus \mathbf{d}$.