

Ekstraktmorfoloogia meetodiga tuletatud keele tehnoloogia vadjä sõnavara näitel

Kristian Kankainen

2019

Sisukord

1	Sissejuhatus	3
2	Teoreetilised lähtekohad	4
2.1	Klassikaline paradigmaatiline morfoloogia	4
2.2	Morfeemi staatus ja definitsioon	4
2.3	Muuttüüp, tüüpsõna ja muutkond	4
3	Ekstraktmorfoloogia meetod	5
3.1	Sissejuhatus	5
4	Vadja morfoloogiliste tüüpsõnade analüüs	7
4.1	Põhivormid ja analoogiavormid	7
4.1.1	Käändsõnad	7
4.1.2	Tegusõnad	7
5	Programmkoodi tuletamine	8
5.1	Keskne kirjeldus Lexical Markup Framework vormingus	8
5.1.1	Sõnaartiklite esitamine LMFis	9
5.1.2	Tüüpsõnamallide esitamine LMFis	9
5.2	Grammatical Framework morfoloogiakomponent	9
5.3	Integreerimine Giella-taristuga	9
5.3.1	Leksikon	10
5.3.2	Paradigmad	10
6	Kokkuvõte	11
7	Põhimõisted ja lühendid	12
		12
8	Kirjandus	13
9	The use of Extract Morphology for Automatic Derivation of Language Technology for Votic	14

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.

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 Klassikaline paradigmaatiline morfoloogia

mh sõna kui selle vormide tervik; pedagoogiline ja üldistuse ülekantavus uuele ainesele Matthews, 1991. Matthewski jätab mudeli vormipõhiseks ja mitte morfeemipõhiseks, selle kohta edasi järgmises allosas.

2.2 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).

2.3 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 kriteeriumitest, ekstraktmorfoloogiat võiks vaadata kui lihtsalt üht väga formaalselt defineeritud muuttüübistikku.

Muutkond sõltub samuti oma alustest, Huldenil on omakorda üks väga formaalselt defineeritud ja ekstraktmorfoloogial põhinev muutkondade süsteem.

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 automaatselt tuletamiseks saadud kirjelduse põhjal. Neid kahte rakendust kirjeldatakse 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. Produktiivsusemudeliga 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
<i>katto</i>	SG NOM
<i>katod</i>	PL NOM
<i>kato</i>	SG GEN
<i>kattoi</i>	PL GEN
<i>kattoje</i>	PL GEN
<i>kattoa</i>	SG PART
<i>kattoi</i>	PL PART
<i>kattoite</i>	PL PART
<i>kattose</i>	SG ILL
<i>kattoise</i>	PL ILL
<i>kattoz</i>	SG INE
<i>kattoiz</i>	PL INE
<i>kattosse</i>	SG ELA
<i>kattoisse</i>	PL ELA
<i>katolle</i>	SG ALL
<i>kattoille</i>	PL ALL
<i>katol</i>	SG ADE
<i>kattoil</i>	PL ADE
<i>katolte</i>	SG ABL
<i>kattoilte</i>	PL ABL
<i>katossi</i>	SG TRAN
<i>kattoissi</i>	PL TRAN
<i>kattossaa</i>	SG TERM
<i>kattoissaa</i>	PL TERM
<i>katoka</i>	SG COM
<i>kattoika</i>	PL COM

Tabel 1: Sisendi muutvormide tabel koos morfoloogiliste tunnustega.

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

Tabel 2: Väljundi tüüpsõnamall (kusjuures $x_1 = kat$ ja $x_2 = o$ vastab 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 Põhivormid ja analoogiavormid

Selles osas selgitatakse välja vadja keele tüüpsõnade põhi- ja analoogiavormid sõnaliigiti.

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.”

4.1.1 Käändsõnad

4.1.2 Tegusõnad

5 Programmkoodi tuletamine

Programmikoodi tuletamise all peetakse siin töös silmas mistahes protsessi, mille käigus tuletatakse mingi üldisema kirjelduse põhjal programmikoodi ühe või mitme konkreetse programmeerimiskeskona jaoks.

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

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

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

Programmikoodi 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 reaalistest 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 ressurssis 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.

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 integreerimine ja paradigmade ehk tüüpsõnamallide integreerimine.

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 Paradigmad

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

Paradigmad esitatakse relatsioonidena sõnavormi ja lemma koos analüüsiga vahel. Sellised relatsioonid sisaldavad lõpmatut hulka sõnalemmasid, millest mõistagi pole suurem osa vajakeelsed. 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 vajja õigekirjakontrollija.

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.

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

Magistritöös esitatud töövoog paneb leksikaalse ressursi keskele 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.

Lemma on suvaliselt valitud grammatiliste tunnuste komplekt, mida kasutatakse lekseemi viitamiseks. 12

Mikrostruktuur on sõnastiku sõnaartikli sisemine struktuur. 12

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. 12

Tehniline tüvi on tähtkoostiste järjend, millega saab tüüpsõnamalli muutvormide muutujad asendada elik väärtustada ja niiviisi koostada ühe konkreetse sõna kõik vormid. 12

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. 5, 12

8 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.
- 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.
- 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.
- 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).
- 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.

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="asTšiuutto"/>
  <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 *tšiuutto* (mille alla kuuluvad mh *tšiuutto* ja *katto*) esitus LMFis. Esitus mudeldab muutvormimalle $x_1 \oplus \mathbf{t} \oplus x_2$ ning $x_1 \oplus x_2 \oplus \mathbf{d}$.