Eesti keele morfoloogiline analüsaator ja korpuse kasutamine tema tegemisel

Heiki-Jaan Kaalep

Kokkuvõte

Artikkel kirjeldab eesti keele morfoloogia-analüsaatorit ja seda, milline oli tekstikorpuse kasutamise mõju nii analüsaatori tegemise protsessile kui valminud programmile. See mõju ei piirdu sõnastikuga, vaid on tunda ka algoritmis ja selle arvuti-realisatsioonis. Analüsaatori loomise algusperioodil ei olnud arvutilingvistilisi formaalseid käsitlusi eesti keele sõnamoodustuse kohta. Programmi loomise ja testimise käigus jõudsime me praktiliselt kasutatava algoritmini, mis võimaldab tuletisi ja liitsõnu analüüsida. Nii morfoloogia-analüsaator kui vastav speller on avalikkusele kasutamiseks kättesaadavad.

1. Sissejuhatus

Personaalarvutite ja arvutivõrkude areng kogu maailmas toob endaga kaasa nõudmise keeletehnoloogia toodete (alates spelleritest) järele selliste keelte jaoks, mille arvutilingvistilised kirjeldused on ebapiisavad. Kui fundamentaaluuringute rahastamiseks pole piisavalt ressursse, siis kas see tähendab, et need keeled jäävadki keeletehnoloogia tagahooviks? Õnneks võib mõnikord kõrgeltkvalifitseeritud lingvistide vähesust kompenseerida baasuuringutes tekstikorpuste kasutamisega. Sellise optimismi aluseks on kogemus morfoloogilise analüsaatori ja spelleri tegemisest, mis seisnes algoritmi tsüklilises kontrollimises ja ümbertegemises, tuginedes korpuste töötlemisel saadud tulemustele.

Aastal 1991, kui morfoloogilise analüsaatori ja spelleri tegemine algas, valitses Eesti arvutilingvistikas järgmine olukord. Elektroonilisel kujul oli olemas lihtsõnade sõnastik, mis sobis suurepäraselt sõnamuutuse käsitlemiseks arvuti abil. Kuid väga vähe oli teada eesti keele reaalsest kasutusest tekstides ja polnud tehtud ühtegi arvutilingvistilist katset käsitleda eesti keele sõnamoodustust, ehkki tuletised ja liitsõnad moodustavad kuni 20% eestikeelsete tekstide sõnedest. Seega tuli viimaste käsitlemise algoritm alles luua, kontrollida selle sobivust ja efektiivsust ning mõnel juhul ka ümber teha.

Morfoloogilise analüsaatori loomise ajendiks oli vajadus spelleri järele. Kuna eesmärgiks oli kommertsprodukt, olid algusest peale olulisel kohal algoritmide realiseerimise ja efektiivsuse probleemid.

Tekstikorpuste kasutamine ei olnud mitte ainult sõnastiku parandamise aluseks (mida antud artikkel käsitleb pinnapealselt), vaid ka algoritmi kujundamise ja realiseerimise aluseks.

Käesolevas artiklis kirjeldatakse ainult morfoloogilist analüsaatorit ESTMORFi. Spellerit ei käsitleta, sest ta erineb ESTMORFist ainult selle poolest, et rahuldutakse esimese lubatava analüüsiga ning ei leita sõna algvormi.

2. Korpus

Analüsaatori loomisel kasutati mitut erinevat korpust: Tartu Ülikooli kirjakeele korpust (TÜKK), BNSi uudiste korpust (BUUK) ja ajalehtede korpust (ALK).

TÜKKi loomine algas 1991. a. sügisel TÜ eesti keele laboris [Hennoste 1996]. Algusest peale otsustati, et selleks, et TÜKKist oleks lingvistidele kasu, peab see olema märgendatud. Kogu TÜKKi kuuluv ajakirjandus, 175 000 sõna, on märgendatud lõikude, lausete, numbrite, lühendite, pärisnimede, otsese kõne, tsitaatide ja mitte-standardse keele osas. Ülejäänud korpuse tekstid märgendati lõikude, lausete ja trükitehniliste võtete (nt. rasvane kiri, kaldkiri) osas.

Tegelikult selgus, et morfoloogilise analüsaatori loomisel ei olnud märgendusest mingit kasu. Märgendusest võiks kasu olla kahel moel:

1. Ta võib lihtsustada analüüsi, võimaldades vahele jätta märgendatud pärisnimed, numbrid, lühendid jms sõned, mida tavaliselt sõnastikes ei kajastata ja mis seetõttu ka analüsaatori sõnastikus puuduvad. See oleks siiski liigne lihtsustus programmi jaoks, mis peab toime tulema suvalise tekstiga.
2. Märgendus võib lihtsustada algoritmi loomist sellega, et pakub märgendatud tekstides õigeid lahendusi. Eesti keele puhul oli siiski palju tulusam lasta morfoloogilisel analüsaatoril analüüsida tavalist teksti ja seejärel filtreerida välja pärisnimed suure algustähe järgi, lühendid ja numbrid aga analüüsita jäänud sõnede hulgast. Järele jäänud analüüsimata sõned olidki siis sõnastiku täiendamise ja algoritmi loomise sisendiks.

Meie kasutasime sellist korpuse versiooni, millest märgendus oli eemaldatud. Sel moel saime programmi tööks loomulikuma keskkonna.

BUUKi loomine algas 1994. a. oktoobris. Ta sisaldab BNSi uudiseid, mida saadetakse tellijatele e-posti teel. Saabuvad kirjad arhiveeritakse automaatselt Aastas suureneb BUUK 3-4 miljoni sõna võrra. Tekstides märgendatakse ainult iga uudise algus ja lõpp.

ALK loomine algas 1993 eesmärgiga jälgida keele muutust ajas. ALK sisaldab erinevaid ajalehti perestroika (1989 ja 1991) ja iseseisvuse ajast (1993, 1995 ja 1996). Praegu on ALKis 4 miljonit sõnet. Tekstide märgendus on eri perioodidel ja lehtedel erinev.

3. ESTMORF, eesti keele morfoloogiline analüsaator

ESTMORF on arvutiprogramm suvalise eestikeelse teksti analüüsimiseks. Teda saab kasutada nt. Interneti kaudu (_____ ja viitasid mööda edasi). ESTMORF on realiseeritud nii, et jooksvas tekstis olevaid sõnesid võrreldakse sõnastikus olevate lekseemide kombinatsioonidega. Võrdlemisel ei kasutata 2-tasandilisi reegleid [Koskenniemi 1983].

ESTMORFi peamised omadused on järgmised:

1. ESTMORF on mõeldud eesti kirjakeele jaoks.
2. Sõnamuutuse käsitlus on täielik, kuni viimase erandini.
3. ESTMORFi sõnastik sisaldab põhisõnavarasse kuuluvaid lihtsõnu ja sagedamaid pärisnimesid ja lühendeid. Produktiivselt moodustatavaid tuletisi ja liitsõnu reeglina sõnastikus pole.
4. Tuletisi ja liitsõnu analüüsitakse algoritmiliselt. Seega pole vaja neid hoida sõnastikus ning on võimalik korrektselt analüüsida ka uusi tuletisi ja liitsõnu
5. Tuletiste ja liitsõnade analüüsi algoritm on koostatud selliselt, et leida iga sõna puhul tema kõige tõenäolisem jaotus komponentideks.
6. Analüüs tugineb sõnastikule ega sisalda heuristikat.
7. ESTMORF hoolitseb ise kirjavahemärkide ja mitmest sõnast koosnevate võõrpärisnimede eest.

8. ESTMORF ei pretendeeri originaalsusele eesti keele morfoloogiasüsteemi käsitlemisel, v.a. sõnamoodustuse osas.
9. Korrektsed analüüsid antakse u. 97% sisendteksti sõnedele. Analüüsimata jäävad haruldased sõnad nagu pärisnimed, lühendid, terminid, släng jms.
10. ESTMORF on morfoloogilise analüüsi vahend, nii teoreetilisteks kui praktilisteks eesmärkideks.
11. ESTMORF ei arvesta süntaktilisi ega semantilisi omadusi nagu valents, transitiivsus või loendatavus.

Arvutimorfoloogias on praegu enam-vähem standardseks meetodiks 2-tasandiliste reeglite kasutamine ja vasakult paremale e. juurest algav sisendsõnede analüüs, vt. nt. [Sproat 1992]. Neid printsiipe on rakendatud nii paljudes erinevates morfoloogilistes analüsaatorites nii paljude eri tüüpi keelte jaoks, et nende loendamine siin läheks liiga pikaks. Kuid ESTMORF ei kasuta 2-tasandilisi reegleid ja sõnesid analüüsitakse paremalt vasakule, s.t. kasutades lõppude ja liidete mahalõikamist. Oma meetodiga kuulub ta analüsaatorite hulka, mida on ka varem aglutinatiivsete keelte jaoks kasutatud, nt vene [Itogi 1983], soome [Brodda ja Karlsson 1980] ja ungari keele [Proszeky ja Tihanyi 1992] jaoks.

Põhjuseks, miks ESTMORF kasutab vanemat meetodit, on tema loomise eesmärk: luua "must kast" morfoloogiliseks analüüsiks ja sõnakontrollimiseks. Sellisest keeletehnoloogilisest vaatepunktist ei ole sõnamuutuse käsitlemise täpne mehhanism tähtis; tähtsad on ainult õiged tulemused. Oli väga mugav teisendada Ülle Viksi "Väike vormisõnastik" (VVS) [Viks 1992] sellisele kujule, mida saab kasutada lõppe maha lõikaval morfoloogilisel analüüsil, ilma et oleks pidanud formuleerima 2-tasandilisi reegleid. Rohkem pöörati ESTMORFi loomisel tähelepanu küsimustele:
1. Kui hästi katab sõnastik reaalsetes tekstides esinevat sõnavara?
2. Kuidas tuleks käsitleda tuletisi ja liiteid?
3. Mida tekstid veel sisaldavad peale tavaliste sõnade ja kuidas neid tuleks käsitleda?

Vastused neile küsimustele määravad lõppude lõpuks morfoloogilise analüsaatori kasulikkuse ja neile vastamine võttis valdava osa ESTMORFi loomisele kulunud ajast.

Algusest peale peeti oluliseks programmi töökiirust, sest täpne töövahend, mis töötab liiga aeglaselt on mittekasutatav. Kuna sisendsõnes pole morfeemide piire näidatud, siis peab analüsaator neid mõistatama ja kontrollima mitmetest loenditest, kuni on leitud vastuvõetav sõna struktuur. Võiks arvata, et paremalt vasakule analüüs on efektiivsem, kuna lõppude ja liidete loendid on lühemad kui tüvede sõnastik ja seega võimaldavad kiiremat otsimist. Seega oleks mõttekas otsida tüve alles siis kui sobiv lõpp on leitud. Samas on teada, et vasakult paremale töötavad analüsaatorid on väga efektiivsed [Karlsson 1992], [Solak ja Oflazer 1993]. Peale hiilgava inseneritöö (kiire otsimine sõnastikust, hea andmete kokkupakkimine, õige programmeerimiskeele ja riistvara valik) võiks arvestada ka reaalsete tekstide statistilisi omadusi ja eriti seda, millise struktuuriga sõnad moodustavad kui suure protsendi tekstist. Sel kohal osutuvadki korpused algoritmide loomisel kasulikuks. ESTMORFi loomise käigus kirjutasime me mõnikord programmi osad ümber, pärast seda kui olime analüüsinud sõnade statistilisi omadusi reaalsetes tekstides.

Alljärgnevalt kirjeldame lühidalt ESTMORFi tööalgoritme. Põhjalikku ülevaadet ESTMORFist vt [Kaalep 1996].

3.1 Lihtsõnade analüüs

Lihtsõnade analüüs on lõpu mahalõikamise, sõnastikust tüve otsimise ja lõpu ning tüve kokkusobivuse kontrollide tsükkel:

Algul võetakse sõna lõpust maha mõned tähed; kontrollitakse, kas need moodustavad eestikeelsele sõnale sobiva lõpu; kui jah, siis kontrollitakse, kas sõna esimene jupp leidub tüvede sõnastikus; kui jah, siis kontrollitakse veel, kas tüvi ja lõpp sobivad kokku. Näiteks sõna ütel[da koosneb tüvest ütel ja lõpust da. Kokkusobivuse kontroll on vajalik selleks, et välja praakida sõnad nagu ütel[ta ja visa[da, kuigi nad mõlemad koosnevad eraldivõetuna normaalsest tüvest ja lõpust

3.2 Tuletiste analüüs

Umbes 8% kõigist sõnedest eestikeelses tekstis on tuletised; ajakirjanduses on neid veelgi rohkem.

ESTMORF kasutab 40 produktiivset järelliidet, mis võivad liituda nimi-, omadus-, arv- või tegusõnale, andes tulemuseks nimi-, omadus- või määrsõna. Mõned järelliited sobivad ainult ühele sõnaliigile, mõned mitmele, andes tulemuseks samuti mitmeid erinevaid sõnaliike. Liitumist kitsendavad piirangud puudutavad tüve sõnaliiki, tüve vormi (nt. nimetava või omastava tüvi) ja tüve lõputähti.

Nt. *dus* võib liituda tegusõna umbisikulise tegumoe mineviku kesksõnale (*töödeldud: töödeldus*) või *eda*-lõpulisele omadussõna ainsuse omastavale, asendades *eda edus*-iga (*müreda: müredus*).

Paljud järelliited võivad kombineeruda. Nt. *ja* ja *lik* annavad *jalik*, nagu *õpetaja*, *õpetajalik*. ESTMORF ei sisalda järelliidete kombineerumise algoritmi, vaid kasutab rohkem kui 100 lubatud kombinatsioonist koosnevat loendit.

Eesti keelt on tavaliselt kirjeldatud kui keelt, millel on väga vähe eesliiteid: ainult *eba* ja *mitte* ning mõned võõrliited, nt. *anti, pro, pseudo* jne. ESTMORF seevastu käsitleb 70 sageli esinevat esikomponenti kui eesti keele eesliiteid, mis võivad liituda nimi-, omadus-, määr- või tegusõnale. Peale selle on veel 30 võõrliidet, mis võivad liituda nimi-, omadus- või tegusõnale.

Eesliidete loendi koostamisel lähtuti järgmistest puhtformaalsetest kriteeriumidest. Liitsõnakomponent tuleks panna eesliidete loendisse, kui:

1. Komponent ei esine omaette sõnana või on tal omaette sõnana selgelt teistsugune tähendus kui liitsõnas, nt *ala* (pind, valdkond) tähendab liitsõnades hoopis *alam-, sub-*.
2. Ei ole silmnähtav, kuidas komponenti moodustada lihtsõnast lähtudes.
3. Komponenti saab vabalt kasutada uute sõnade moodustamiseks
4. Komponent esineb tekstides küllalt paljudes sõnades.

ESTMORF on küllaltki range ja kahtlase reegli formuleerimise asemel hoitakse paljusid tuletisi sõnastikus. Nt. eesliide *nüüdis-* võib liituda nimisõnadele, nt. *nüüdisooper*, kuid mitte omadussõnadele, nt. **nüüdislai*. Erandlik omadussõna *nüüdisaegne* on pandud sõnastikku.

3.3 Liitsõnamoodustus

Liitsõnamoodustus on eesti keeles isegi produktiivsem nähtus kui tuletus. Liitsõnu on eestikeelsetes tekstides keskmiselt 12%; ajalehetekstides veel rohkem.

Reeglid ja piirangud, mis liitsõnade moodustamist määravad, võib jagada kahte suurde gruppi:

1. Liitsõna komponentide arv
2. Komponentide eneste omadused: nt. kas komponent on tüvi või järelliide; mis sõnaliiki tüvi kuulub, millised tähed on tüve lõpus jne

Liitsõnade moodustamisest võivad põhimõtteliselt osa võtta järgmised 8 lihtstruktuuri: tüvi, tüvi + lõpp, tüvi + järelliide, tüvi + järelliide + lõpp, eesliide + tüvi, eesliide + tüvi + lõpp, eesliide + tüvi + järelliide, eesliide + tüvi + järelliide + lõpp

Teoreetiliselt võiksid nad omavahel kombineeruda kuidas tahes, kuid reaalsetes tekstides on sagedasemate mallide pingerida järgmine:

Liitsõna-mall	**% kõigist liitsõnadest**
tüvi + tüvi	70-75%
tüvi + tüvi + järelliide	5-10%
tüvi + tüvi + tüvi	5-10%
tüvi + lõpp + tüvi	1-5%
tüvi + lõpp + tüvi + järelliide	1-5%
tüvi + järelliide + tüvi	1-5%

On terve hulk nõudeid, millele iga malli komponendid peavad vastama. Need nõuded on väga sarnased piirangutega, mida kasutatakse tuletiste puhul ja nad puudutavad tüve sõnaliiki, tüve vormi ja tüve lõputähti. ESTMORF võtab arvesse ainult formaalseid piiranguid; liitsõna tähenduslikku sobivust ta ei arvesta.

ESTMORF kasutab ka kahte tüvede loendit, mis võivad osaleda liitsõnade moodustamisel vabamalt kui muud tüved: tõenäolisemate esi- ja järelkomponentide loendeid.

Liitsõna tükeldamisel osasõnadeks on sageli võimalik mitu varianti, nt. lae+kaunistus ja laeka+unistus. ESTMORF leiab ainult ühe liitsõna tükeldamise variandi. Liitsõnade analüüs on alamprogrammide järjekorra ja sõnaloendite valiku abil organiseeritud sel moel, et väljundiks oleks kõige tõenäolisem analüüs, antud näite puhul lae+kaunistus. Peamiseks juhiseks seejuures on põhimõte, et komponentide arv peab olema minimaalne: eelistada tuleb lihtsõnu tuletistele ja liitsõnadele ning vähema komponentide arvuga liitsõnu keerulisematele.

Pärast mitmeid katsetusi oleme jõudnud järgmise variantide proovimise järjekorrani, mis annab vähima vigade arvu. Algoritm ei ole puhtalt vasakult paremale ega paremalt vasakule suunatud analüüs.

1. Kas sõna on lihtsõna?
2. Kas sõna on struktuuriga tüvi + järelliide (või tüvi + järelkomponent)?
3. Kas sõna on struktuuriga eesliide + tüvi (või esikomponent + tüvi)?
4. Kas sõna on struktuuriga tüvi + tüvi?
5. Kas sõna on struktuuriga tüvi + tüvi + järelliide (või tüvi + tüvi + järelkomponent)?

6. Kas sõna on struktuuriga eesliide + tüvi + järelliide (või esikomponent + tüvi + järelliide või eesliide + tüvi + järelkomponent või esikomponent + tüvi + järelkomponent)?
7. Kas sõna on struktuuriga tüvi + tüvi + tüvi?
8. Kas sõna on struktuuriga tüvi + lõpp + tüvi?
9. Kas sõna on struktuuriga tüvi + lõpp + tüvi + järelliide (või tüvi + lõpp + tüvi + järelkomponent)?
10. Kas sõna on struktuuriga tüvi + järelliide + tüvi (või tüvi + järelliide + tüvi + järelliide või tüvi + järelliide + tüvi + järelkomponent)?
11. Kas sõna on struktuuriga eesliide + järelkomponent (või esikomponent + järelkomponent)?
12. Kas sõna on struktuuriga eesliide + liitsõna (või tüvi + liitsõna)?

4. Morfoloogiline mitmesus

Morfoloogiline analüsaator annab analüüsitava sõna puhul sageli mitu võimalikku analüüsivarianti. Sõnavorm võib olla mitteüheselt tõlgendatav kahel põhjusel:
1. On mitu viisi jagada sõna lekseemideks, nagu nt.

kapsas
  kapsas+0 //_S_ sg n, //
  kapsas+s //_S_ sg in, //
  kapsa+s //_V_ s, //

2. Lekseemid on samad, kuid neid saab tõlgendada erinevalt, nagu nt:

lisasid
  lisa+sid //_S_ pl p, //
  lisa+sid //_V_ sid, //

Mõnikord võib lõpuformatiivi tõlgendada sõna paradigma siseselt mitmel moel, nagu nt. *sid* ülaltoodud näites tegusõna *lisasid* puhul (teine või kolmas pööre). Kui lõpuformatiiv on mitme vormi osas homonüümne kõigi antud sõnaliigi sõnade puhul, nagu *sid* tegusõnade puhul, siis võib antud formatiivi osas morfoloogilisi kategooriaid ühendada. Seda teeb tegusõnade osas VVSi [Viks 1992] eeskujul ka ESTMORF.

ESTMORFi abil analüüsitud tekstidest nähtub, et üle 42% sõnadest TÜKKis on morfoloogiliselt mitmesed. See on flektiivse keele jaoks suur arv. Tegelikult on eesti keeles morfoloogilist mitteühesust isegi rohkem, arvestades seda, et tegusõnadel on homonüümsete formatiivide puhul morfoloogilised kategooriad ühendatud ja seda, et liitsõnade analüüsil ei väljastata kõiki võimalikke komponentide kombinatsioone, vaid ainult kõige tõenäolisemad.

5. ESTMORFi sõnastik

Hea sõnastiku loomine on kõige ilmsem kasu, mida võib saada korpuse kasutamisest. ESTMORFi sõnastikus on 38 000 sõna. Ta põhineb VVSi elektroonilisel versioonil [Viks 1992]. Kuna kõik tüvevariandid on sõnastikus eraldi sees, siis on seal 67 000 tüve. Võrreldes ESTMORFi sõnastikku VVSiga näeme, et sinna on lisatud mitutuhat sõna:
1. Ligikaudu 1200 põhisõnavarasse kuuluvat lihtsõna
2. Ligikaudu 2500 liitsõna, mille moodustamine on algoritmiliseks kirjeldamiseks liiga keeruline või ebaregulaarne. Need 2500 sõna esindavad järgmisi sõnaliike: 100 tegusõna, 870 määrsõna, 150 arvsõna, 8 asesõna, 1300 nimi- ja omadussõna.

3. Ligikaudu 2700 pärisnime ja 500 genitiivatribuuti, s.h. u. 70 võõrpärisnime, mis koosnevad mitmest sõnast nagu *New York*.
4. Ligikaudu 300 lühendit.

VVSist on eemaldatud mitutuhat sõna:
1. Ligikaudu 1800 vananenud või murdesõna
2. Ligikaudu 2700 liigset tuletist (VVS sisaldab palju produktiivseid tuletisi)

6. ESTMORFi loomise etapid

Alljärgnevalt kirjeldame ESTMORFi arengu ja testimise etappe. Igal etapil leiti ja parandati algoritmi, sõnastiku ja programmi vigu. Ka ESTMORFi ja tema alusel loodud spelleri kasutajad aitasid arendamisele kaasa, tuues välja ESTMORFi vigu.

ESTMORFi loomine algas augustis 1991, kui saime VVSi [Viks 1992] elektroonilise versiooni. VVS sisaldab u. 36 000 lihtsõna koos täieliku kirjeldusega, et genereerida sõna paradigma kõik vormid. Lihtsõnade analüsaator loodi 4 kuuga, augustist detsembrini 1991.

Analüsaator suutis analüüsida 75% sisendteksti sõnedest. See oli tuletiste ja liitsõnade analüüsi algoritmi loomise lähtekoht. Me ei teadnud vastust küsimustele:
1. Kui produktiivne on sõnamoodustus reaalsetes tekstides?
2. Millised on tuletuse ja liitsõnade moodustamise mallid ja millised neist on produktiivsed?

Varasemad uurimused ([Kask 1967], [Kull 1967], [Kasik 1984] ja [Kasik 1992]) andsid kasulikke vihjeid mõlemale küsimusele vastamiseks, aga ei olnud koheselt kasutatavad. Lisaraskusi tekitas asjaolu, et sõnamoodustust oli kirjeldatud kui sünteesiprotsessi, meid huvitas aga analüüs. Eraldi probleemiks oli veel see, et liitsõnade moodustust oli kirjeldatud kui kahe komponendi liitmist, samas kui reaalsetes tekstides esineb kuni 5-komponendilisi liitsõnu. Ei olnud selge, kuivõrd võib keerulisema struktuuriga sõnade puhul kasutada rekursiivselt samu reegleid, mida kasutatakse 2-komponendiliste sõnade puhul.

Et lihtsustada liitsõnade analüüsi algoritmi väljatöötamist, jagasime ülesande kaheks:
1. Leida, millise struktuuriga liitsõnu reaalsetes tekstides kasutatakse
2. Leida, millised on piirangud iga struktuuri kasutamisele

Uurisime kõiki struktuure eraldi ja püüdsime piirangute formuleerimisel olla väga ranged, nii et lubatud oleksid ainult kindlasti õiged kombinatsioonid. Vaatame nt. struktuure tüvi1+tüvi2 ja tüvi1+tüvi2+tüvi3. Algul me eeldasime, et tüvi1 ja tüvi2 võivad 2-komponendilisse struktuuri kombineeruda vabamalt kui 3-komponendilisse, sest tekstides on 3-komponendilisi liitsõnu vähem kui 2-komponendilisi. Kui me hiljem leidsime, et oleme olnud liiga ranged, siis lõdvendasime piiranguid. Iga kord kui muutsime analüüsitavate struktuuride ja piirangute hulka, kontrollisime tulemusi samade tekstide peal mida olime varem kasutanud. Tundmatute ja valesti analüüsitud sõnade hulk vähenes samm-sammult. Kui see oli küllalt väike, võtsime me testimiseks uued tekstid ja kogu tsükkel kordus.

Sõnamoodustuse algoritmi parandamise lõpetasime siis, kui jõudsime punkti, kus:
1. Uues tekstis jäi analüüsimata sama palju tuletisi ja liitsõnu kui jäi analüüsimata lihtsõnu.
2. Need lihtsõnad olid nii erandlikud (nt. släng, murre, tehnilised terminid), et neid ei tohiks analüsaatori sõnastikku lisada.

Järeldasime, et olukord oli sarnane lihtsõnade analüüsile: ilmselt esindasid need tuletised ja liitsõnad selliseid haruldasi malle eesti keele sõnamoodustuses, mida võib lugeda erandlikeks või mitte-grammatilisteks ja mida seetõttu ei tohiks algoritmis kajastada.

Peale äratuntavate sõnade hulga maksimiseerimise pidasime silmas ka algoritmi kiirust. Et minimiseerida aega, mis kulub katsetele analüüsida sisendsõnesid vale struktuurimalli alusel, korraldasime me programmi töö nii, et algul proovitakse tõenäolisemaid struktuure.

Kui liitsõna esindas harvaesinevat malli, siis oli lihtsam ta tervikuna sõnastikku lisada kui algoritmi erilisel moel teisendada.

Sõnamoodustuse algoritmi väljatöötamise esimene etapp algas jaanuaris 1992 ja kestis 1994. aastani. Selle tulemusena loodi speller ja anti sõltumatutele kasutajatele. Kasutatud korpus oli väike: ainult 100 000 sõna ilukirjandustekste TÜKKist ja mitmed majasisesed tekstid (artiklid, kirjad jms). Iga korpuses ette tulevat mitteanalüüsitud sõna uuriti hoolikalt, sõltumata tema esinemissagedusest, et otsustada kui tüüpiline ja loomulik ta on. Algoritmi muudeti ainult juhul, kui leiti et sõna esindab piisavalt tüüpilist juhtumit.

Algul eeldas ESTMORF, et ainult 2-komponendilised liitsõnad on vabalt lubatavad. Et saaks analüüsida keerulisemaid liitsõnu, toodi sisse kaks loendit: tüved, mis võivad liituda sõna algusse ja tüved, mis võivad liituda sõna lõppu. Sõnastikku lisati mitutuhat ebaregulaarset liitsõna: sellist, mille mõni komponent ei esine üksiksõnana või kuulub sõnaklassi, mis tavaliselt sõnamoodustuses ei osale. Loendite ja ebaregulaarsete liitsõnade allikaks olid VVSi lisad 2-4 [Viks 1992]. Liitsõnade analüüsi algoritm ei teinud selget vahet eri struktuuridel. Nt. oleks mõistlik tundmatu liitsõna analüüsil algul kontrollida kõiki võimalusi, et sõna struktuur on tüvi1+tüvi2 ja alles ebaõnnestumise korral kontrollida, kas struktuur on tüvi1+tüvi2+tüvi3. ESTMORF aga püüdis leida mingit komponenti sõna algusest ja seejärel iga hinna eest analüüsida ülejäänud osa sõnast, katsetades kõikvõimalikke tüvede, liidete ja lõppude kombinatsioone.

1994 lugesime ESTMORFi piisavalt valmis olevaks, et kasutada teda kvantitatiivse lingvistika töövahendina. Me analüüsisime 300 000 sõnalist ilu- ja ajakirjanduse alamkorpust TÜKKist ja 100 000 sõnalist ajakirjanduse alamkorpust ALKist aastaist 1989 ja 1991. Selgus, et ESTMORF ei tundnud 4% sõnu ilukirjandusest ja 9% ajakirjandusest, peamiselt pärisnimesid, lühendeid ja numbreid sisaldavaid sõnesid. Seega tuli lisada ESTMORFi sõnastikku lisada tuhandeid pärisnimesid ning algoritm numbreid sisaldavate sõnede ja muude tekstis esinevate mitte-sõnade analüüsiks.

Selgus ka, et algoritm ei olnud reaalselt tekstis ette tulevate sõnade struktuuride suhtes optimaalne. ESTMORF alustas pikima võimaliku lõpu mahalõikamisest, eeldades, et kui lõpp on lubatav lõpp, siis on väga tõenäoline, et ülejäänud osa sõnast on lubatav tüvi ja seega saab minimiseerida sõnastiku poole pöördumiste arvu. Testimisel leidsime aga, et enam kui pooled tekstis esinevad sõnavormid on kas muutumatud sõnad või null-lõpuga, nt. *raha*, ning et lihtsõnad moodustavad 75-85% eestikeelsest tekstist. Seega kui alustada analüüsi sellest, et lihtsalt kontrollida sõne olemasolu sõnastikus, siis 40% juhtudest saame positiivse vastuse juba esimesel katsel. Testimise andmetele tuginedes muutsime lihtsõnade analüüsi algoritmi. ESTMORF alustab nüüd lühima võimaliku lõpu mahalõikamisest, sest mida lühem on lõpp, seda tõenäolisem on tema esinemine. Katsed erinevate lõpu mahalõikamise strateegiatega näitasid, et lühematest alustamine andis tulemuseks kiirema spelleri.

Liitsõnade analüüsis eraldasime eri struktuure töötlevad moodulid ja järjestasime nad nii, et tõenäolisemaid struktuurimalle proovitakse enne.

1995 testisime ESTMORFi ühe kuu uudiste e. 500 000 sõna peal BUUKist. Selle tulemusena lisati sõnastikku veelgi pärisnimesid. Selgus ka, et tekstid sisaldasid palju kirjavigu. 800-sõnalist vigaste sõnade loendit kasutati ESTMORFi parandamiseks: imiteerides spelleri soovitusmoodulit, genereeriti vigastest sõnadest uued sõnad ja kontrolliti nende korrektsust. ESTMORF ei olnud piisavalt range veidrate sõnade väljapraakimisel. Seega otsustasime sisse tuua veel ühe loendi: selliste tüvede loendi, mis ei saa esineda liitsõna komponendina. Iga kord kui ESTMORF leiab liitsõna analüüsides ühe võimaliku komponendi, kontrollib ta, kas see pole komponentide "mustas nimekirjas".

1996 lemmatiseerisime EV Valitsuse 1995. a. otsuste ja määruste 300 000 sõnalise korpuse. Selle tulemusena lisati sõnastikku veelgi (peamiselt vene) pärisnimesid. Analüüsisime ka kogu TÜKKi (1 miljon sõna). Selgus, et ESTMORFi 32 000 sõnalisest lihtsõnade sõnastikust ei esinenud korpuses kordagi 15 000 sõna. See 15 000 sõnaline loend kontrolliti käsitsi läbi ja eemaldati 1800 vananenud ja murdesõna.

1996. aastal näitas G. Orwelli "1984" (79 000 sõna) analüüs, et ESTMORFi võib lugeda enam-vähem lõpetatuks. Ainult 2% sõnadest, peamiselt Uuskeele sõnad ja briti pärisnimed, jäi analüüsimata.

7. Kokkuvõte

Eesti keele morfoloogilise analüsaatori ja spelleri loomisel oli eesmärgiks luua programm, mis suudaks analüüsida tavalist teksti ilma kunstlike piiranguteta. Eesti keele sõnamoodustuse arvutilingvistilise kirjelduse puudumine raskendas võetud ülesannet. Pärast viit aastat tööd programmi kallal võime öelda, et oleme oma eesmärgi saavutanud.

Programmi loomine oli iteratiivne protsess: algul loodi programm, siis testiti teda korpuse peal, seejärel analüüsiti tulemusi ja muudeti programme. Siis tsükkel kordus.

Testimisel ja korpuse analüüsil kasutatavad meetodid olid väga lihtsad. Me ei läinud kaugemale sagedusloenditest, protsentidest ja lihtsast väljundite võrdlemisest.

Loomulikult oli korpuse kasutamisel suur mõju programmi poolt kasutatavale sõnastikule. Kuid see võimaldas meil ka leida küllalt hea algoritmi produktiivse sõnamoodustuse jaoks. Lisaks sellele tõi korpusel testimine kaasa ka muutused lihtsõnade analüüsi algoritmis.

8. Tänuavaldused

ESTMORF poleks olnud võimalik ilma Ülle Viksi "Väikese vormisõnastiku" elektroonilise versioonita. Paljud ESTMORFi olulised moodulid programmeeris Tarmo Vaino. Viire Villandi valis ja sisestas suure hulga pärisnimesid ning kasutas ESTMORFi tema testimisfaasis. Toomas Mattson, Ülle Viks, Heili Orav, Kadri Muischnek ja Microsoft WPG kasutasid, testisid ja andsid väärtuslikke nõuandeid ESTMORFi ja tema alusel loodud spelleri kohta. TÜ Üldkeeleteaduse õppetooli poolt olid kõik kasutatavad korpused.

9. Kirjandus

Brodda and Karlsson 1980 - Brodda, B. and Karlsson, F. “An Experiment with Automatic Morphological Analysis of Finnish.” *Papers from the Institute of Linguistics. Publication 40*, Stockholm: University of Stockholm, 1980.

Hennoste 1996 - Tiit Hennoste, “Tartu University Corpus of Written Estonian: a survey of the structure of texts and principles of selection.” Kogumikus *Estonian in the Changing World* , toim. Haldur Õim, Tartu 1996. lk. 7-32.

Itogi 1983 - *VINITI Itogi nauki i tehniki. Serija informatika*, Tom. 7. Moskva, 1985

Kaalep 1996 – Heiki-Jaan Kaalep, “ESTMORF: A Morphological Analyzer for Estonian” Kogumikus *Estonian in the Changing World* , toim. Haldur Õim, Tartu 1996. lk. 43-98.

Karlsson 1992 - Karlsson, F. “SWETWOL: a Comprehensive Morphological Analyzer for Swedish.” *Nordic Journal of Linguistics* 1, (1992), 1-45.

Kasik 1984 - R. Kasik. *Eesti keele tuletus›petus: õppevahend eesti filoloogia ja zurnalistikaosakonna üliõpilastele. 1. Substantiivituletus.* TRÜ, Tartu 1984

Kasik 1992 - R. Kasik. *Eesti keele tuletusõpetus: õppevahend eesti filoloogia ja zurnalistikaosakonna üliõpilastele. 1. Adjektiivi- ja adverbituletus.* TRÜ, Tartu 1992

Kask 1967 - Kask, A. “Liitsõnad ja liitmisviisid eesti keeles.” *Eesti keele grammatika 3.1*., Tartu, 1967

Koskenniemi 1983 - Koskenniemi, K. “Two-level Morphology: A General Computational Model for Wordform Recognition and Production.” *Publications of the Dept. Of General Linguistics, University of Helsinki*, 11 (1983)

Kull 1967 – Kull, R. *Liitnimisõnade kujunemine eesti kirjakeeles.* Dissertation for candidate of philological sciences, ENSV TA KKI, Tallinn 1967

Proszeky and Tihanyi 1992 - Proszeky, G. and Tihanyi, L. “A fast Morphological Analyzer for Lemmatizing Agglutinative Languages.” Kiefer, F. G. Kiss and J. Pajzs (Szerk.) *Papers in Computational Lexicography. Complex-92*, Budapest: Linguistics Institute, HAS, 1992, pp. 265-278

Solak and Oflazer 1993 - A. Solak and K. Oflazer, “Design and Implementation of a Spelling Checker for Turkish.” *Literary and Linguistic Computing*, Vol. 8, No. 3, 1993

Sproat 1992 – R. Sproat, *Morphology and Computation.* The MIT Press, Cambridge, Mass.

Viks 1992 - Ü. Viks *A Concise Morphological Dictionary of Estonian.* Tallinn: Institute of Estonian Language and Literature, 1992