Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis

Heiki-Jaan Kaalep, Tarmo Vaino
Tartu Ülikool

1. Lingvisti töövahendite komplekt

Oma töös kasutab lingvist mitmesugust keelematerjali: nii tektsikorpusi e. -kogusid kui sõnastikke. Meie kirjeldame töövahendeid, mis on sellise lingvisti käsutuses, kes kasutab elektroonilisel kujul olevat tekstilist materjali.

Viimast on praegu võimalik kätte saada rohkem kui kunagi varem. Aga mida hakata peale tekstidega, mida on nii palju, et neid ei jõua kuidagi läbi lugeda? Lahendus on ilmne: lingvist peab teda huvitava materjali tekstimassist kuidagi välja filtreerima. Hea oleks materjali seejuures ka sellisel viisil esitada (grupeerida), et käsitsi ülevaatamine, millest ju nagunii pääsu pole, oleks lihtsam.

Seega vajab lingvist tarkvara. Samas on üks huvipakkuv lingvistiline probleem tüüpiliselt niivõrd mitte-standardne, et valmis arvutiprogrammi tema lahendamiseks pole olemas. Võib isegi öelda, et valmis programm saabki olemas olla ainult ebahuvitava probleemi käsitlemiseks. Nüüd on lingvistil justkui kaks võimalust: paluda IT-asjatundjat, et see talle vajaliku programmi kirjutaks, või kirjutada ise. Kumbki tee ei ole ahvatlev. Esimesel juhul on loomulik, et tekivad kommunikatsiooniraskused täpseid formuleeringuid nõudva IT-spetsialisti ja lingvisti vahel, kes, lahendades lingvistilist probleemi, ei tea ju ka ise, kuidas seda teha. Tulemuseks on, et IT-spetsialist peab programmi aina ümber tegema… Kui aga lingvist ise asub programmi kirjutama, siis kulub tal selleks reeglina palju rohkem aega kui IT-spetsialistil, kusjuures tema oma spetsiifilised oskused jäävad seejuures rakendamata.

Meie meelest on heaks lahendusteeks tarkvara-lego lähenemisviis. Selle mõte on, et konkreetset ülesannet saab lahendada, kombineerides väikest hulka standardprogramme, nagu lego klotsidest ehitatakse keerulisi süsteeme.

Me ei püüa siin anda ammendavat loendit lingvistile vajalikest töövahenditest, mõned lihtsamad toome aga ära küll. Seejuures tugineme enda ja kolleegide mitme-aastasele kogemusele, mis on just sageli kasutatavad töövahendid välja setitanud.

Keskkonnaks, mida me kasutame, on UNIX. Selle põhjuseks on see, et esiteks on UNIXis lai valik tekstide töötlemiseeks sobivaid käske, teiseks see, et neid käske saab väga lihtsaslt omavahel kombineerida, ning lõpuks see, et UNIXi keskkond on stabiilne: inimene, kes õppis UNIXit kasutama aastal 1980, saab oma oskusi kasutada ka aastal 2000 ja ilmselt edaspidigi; kes õppis aga DOSi või Windowsi kasutama, peab iga paari aasta tagant ümber õppima. On selge, et pidev ümberõppimine takistb süvenemist.

Niisiis, käsud ("klotsid"), millest lingvistile vajalikke filtreid koostatakse.

1. grep. Selle abil saab tekstist välja võtta kõik meid huvitavat sõna, väljendit või nende kombinatsiooni sisaldavad read. Nt. sõnastikust kõik ma-lõpulised sõnad, mis pole tegusõnad, või kõik read korpusest, kus esineb "poole rohkem".
2. sed. Selle abil saab ridu teisendada. Nt. uurides autorikõnet võib eemaldada kõigist ridadest teksi, mis on jutumärkide vahel.
3. tr. Selle abil saab mugavamalt kui sed-iga üksikuid tähti teisendada, nt. suuri väikesteks.

4. sort. Selle abil saab ridu järjestada.

5. head. Selle abil saab välja võtta meid huvitavat arvu ridasid faili algusest.

6. tail. Selle abil saab välja võtta meid huvitavat arvu ridasid faili lõpust.

7. paste. Selle abil saab ridu kokku panna nagu tabeli veerge.

8. join. Selle abil saab järjestatud ridadest koosnevaid tabeleid kokku panna (nagu relatsioonilises andmebaasis pannakse võtme järgi kokku tabeli veerge). On mugav kasutada nt erinevate sõnastike kombineerimiseks.

Kõiki käske saab kombineerida, nii et ühe käsu täitmisel saadav tulemus on teise sisendiks, ilma et peaks vahepeal midagi faili kirjutama. Kuidas UNIXi käske täpselt kasutada ning kombineerida, selleks on olemas küllaldaselt õpikuid ning UNIXi enda dokumentatsiooni, mida siinkohal ei loetle.

Ülalkirjeldatud UNIXi käskude kasutamist lingvistile vajalikul moel on ehk kõige paremini kirjeldanud Ken Church oma käsikirjas "Unix for Poets" (Church), kelle tööst oleme inspiratsiooni saanud.

On selge, et kuigi standardsed töövahendid on head, oleks siiski vaja ka spetsiaalselt lingvistidele mõeldud käske nt. lausepiiride leidmiseks või KWIC-ideksite esitamiseks. On hea, kui neid saab kasutada samasuguste ehitusklotsidena nagu UNIXi enda käske.

Eesti keele puhul on vaieldamatult vajalikuks "ehitusklotsiks" morfoloogiline analüsaator, s.t. programm, mis suvalises vormis sõna puhul tekstis võib määrata selle sõna algvormi, sõna struktuuri (formatiivid) ja morfoloogilise informatsiooni (nt. sõnaliigi, käände või pöörde, arvu jms).

Üks tüüpiline ülesannete jada, mida lingvist kasutab tekstist teda huvitava materjali väljavõtmisel, ongi nt järgmine:

tekst -> lausepiiride leidmine -> morf. analüüs -> grupeerimine, järjestamine jms

2. Teksti täielik morfoloogiline analüüs

Artiklis kirjeldame lähemalt üht olulist töövahendit eesti keele uurimisel - teksti täielikku morfoloogilist analüsaatorit. Tegemist on programmiga, mille sisendiks on tekst ja väljundiks morfoloogiliselt analüüsitud sõnad, kusjuures ta omistab igale sõnale just antud kontekstis sobiva(d) analüüsivariandi(d). Ideaaljuhul oleks variante üks, kuid programmi praegune variant seda täies ulatuses ei võimalda.

Programm, mida kirjeldame, on mõeldud just nimelt lingvisti töövahendiks, mitte teoreetiliste printsiipide kontrolliks ega illustratsiooniks. Sellest ka tema orienteeritus nn. reaalsete tekstide töötlemisele, mitte hoolikalt valitud sõnade hulgale (nt sõnastikule). Reaalne tekst sisaldab elemente, mida ükski sõnastik ei esita: pärisnimesid, kirjavigu, võõrkeelseid tsitaate, valemeid, neologisme, arhaisme, slängi, murdeid jne. Korralik töövahend peaks suutma neid kuidagi tõlgendada, ideaaljuhul andma nende kõigi korrektse, konkreetsesse konteksti sobiva analüüsi.

Traditsiooniliselt peetakse morfoloogiliseks analüsaatoriks programmi, mis üksikule sõnavormile leiab analüüsi. Nt:

```
Mees
   mees+0 //_S_ sg n, //
   mesi+s //_S_ sg in, //
peeti
   peet+0 //_S_ adt, sg p, //
   pida+ti //_V_ ti, //
```

kinni

kinni+0 //_D_ //

Sellist analüüside paljusust konkreetses tekstis nähes on meie esimeseks intuitiivseks reaktsiooniks, et siin on midagi viltu - inimesele ei tule konteksti mittesobivad analüüsivariandid pähegi, mistõttu arvuti näib pakkuvat meile liiga palju müra. Et tulla vastu inimese intuitsioonile, aga ka mitmete praktiliste (arvuti)lingvistiliste ja keeletehnoloogiliste vajaduste tõttu, on mõttekas teostada morfoloogilist analüüsi konteksti arvestades, nii et väljundis oleks kõik sõnad üheselt analüüsitud.

Teksti täielik morfoloogiline analüüs koosneb kahest etapist: üksiksõnade morfoloogiline analüüs ning ühestamine. Üksiksõnade morfoloogiline analüüs on eesti keele puhul teksti täieliku morfoloogilise analüüsi tingimata vajalik osa (morfoloogiliselt lihtsama keele, nt inglise keele puhul, võib ta ka puududa). Ta annab igale sõnale hulga analüüsivariante. Seejärel toimub mitmest variandist ühe, antud konteksti sobiva valimine e. ühestamine. Meie poolt kirjeldatav ühestaja eeldab, et sõnu vaadeldakse lause kontekstis; laiemat konteksti ei vaadata. Seega ühestamine eeldab, et lausepiirid on juba leitud. Seetõttu ongi vajalik ka programm, mis sisendteksti enne morfoloogilist analüüsi lauseteks jagab - lausestaja.

3. Lausestaja

Lausepiiride leidmine võib tunduda triviaalse ülesandena, kuid seda ta siiski pole. Nt. punkt numbri, initsiaali või lühendi taga võib, aga ei pruugi tähistada lause lõppu. Samuti võivad lauselõpupunktile järgneda sulud, jutumärgid või veel mingid muud sümbolid, mistõttu lause lõpp tuleb alles pärast punkti.

Kuna lausepiiride leidmine on tihti kasutatav ja standardne ülesanne, siis on mõistlik eraldada ta omaette "lego klotsiks", mida vajadusel teiste moodulitega kombineerida.

Lausestaja iseseisva moodulina pakub arvatavasti vähe huvi, kui lingvist tegeleb sõnavaraga. Kui ta tegeleb aga süntaksiga, siis on lausestamine tingimata vajalik etapp. Samuti on ta ilmselt vajalik, kui on vaja leida näitelauseid.

4. Sõnastikupõhine morfoloogiline analüüs

Et teha teksti morfoloogilist analüüsi, kasutatakse tavaliselt sõnavormide töötlemist ja võrdlemist antud keele leksikoniga ning mitmesuguseid heuristilisi reegleid sõnade jaoks, mida leksikonis pole.

Ligikaudu 98±1% eestikeelse sisendteksti sõnadest on analüüsitav sel moel, et kasutatakse sõnastikust järelevaatamist, mitmesuguste morfeemide loendeid ja nende kombineerimise eeskirju. See protsent on suurem kui inglise keele puhul, kus ta on ligikaudu 95 (Voutilainen jt. 1992). Eesti keele morfoloogiline analüüs on realiseeritud nii, et jooksvas tekstis olevaid sõnesid võrreldakse sõnastikus olevate lekseemide kombinatsioonidega. Võrdlemisel ei kasutata 2-tasemelisi reegleid (Koskenniemi 1983) ning sõnesid analüüsitakse paremalt vasakule, s.t. kasutades lõppude ja liidete mahalõikamist ning tüve(de) kontrollimist leksikonist, milles on 38 000 sõna tüved (67 000 tükki).

Sellise analüüsi peamised omadused on järgmised:

1. Ta on mõeldud eesti kirjakeele jaoks.
2. Sõnamuutuse käsitlus on täielik; analüüsitakse ka erandlikke vorme.

3. Analüsaatori sõnastik sisaldab põhisõnavarasse kuuluvaid lihtsõnu ja sagedamaid pärisnimesid ning 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. Korrektsed analüüsid antakse u. 98% sisendteksti sõnedele. Analüüsimata jäävad haruldased sõnad nagu pärisnimed, lühendid, terminid, släng jms.
8. Analüsaator hoolitseb ise kirjavahemärkide ja mitmest sõnast koosnevate võõrpärisnimede analüüsi eest.
9. Ei pretendeerita originaalsusele eesti keele morfoloogiasüsteemi käsitlemisel, v.a. sõnamoodustuse osas.
10. Analüsaator ei arvesta süntaktilisi ega semantilisi omadusi nagu valents, transitiivsus või loendatavus.
11. Analüsaator on aluseks kommertsiaalsele eesti keele spellerile.
Detailset kirjeldust vt. (Kaalep 1997), (Kaalep 1998)

5. Oletaja

Kuni 3% tekstist moodustavad sõnad, mille analüüsimiseks ei ole sõnastikust abi, sest sõnastikust vastavad kirjed puuduvad. See protsent on eri tekstiklassides väga erinev. Suurim, 3% ümber, on ta ajakirjanduse ja informatsiooniliste ning teatme-materjalide puhul; samas kui ilukirjanduse ning seadusetekstide puhul on ta sageli kõigest 0,5.
Ajakirjandustekstide puhul jaguneb see 3% omakorda järgmiselt: ligikaudu 66% tundmatutest sõnavormidest on pärisnimed; 10% üldnimisõnad; 9% ebastandardselt esitatud kirjavahemärgid (nt. mõttekriips); 8% lühendid; 1% mitmesugused numbrikombinatsioonid; 1% omadussõnad, tegusõnad, määrsõnad; 5% võõrkeelsed sõnad, WWW-aadressid jm sümbolijadad, millele on raske üldse mingit mõistlikku analüüsi pakkuda.
Meie programm oletab sõna algvormi ja seda, millises vormis ta on, ainult sõnavormi enda alusel. Arvesse võetakse sõna lõputähti ja silpide arvu. Oletamisel ei arvestata sõna konteksti.
Oletamisel kontrollitakse, kas sõna võiks olla:
1. Lühend (kuni 2 tähte või ilma vokaalideta "sõna"; suurtähtedest koosnev sõna, millele võib olla lisatud väiketäheline käändelõpp).
2. Ilmse kirjaveaga, mille parandamisel on sõna sõnastikku kasutades analüüsitav (nt. sõnadevaheline tühik on jäänud puudu või on kolm ühesugust vokaali kõrvuti).
3. Pärisnimi.
4. Tuletatud sõna või liitsõna, mille puhul on kas kasutatud harvaesinevat moodustusmalli või mis sisaldab sõnastikust puuduvat lihtsõna.
5. Tundmatu lihtsõna: nimisõna või verb (otsustame sõna lõpu ja sellele eelnevate tähtede ning silpide arvu alusel).
Oluline abi on oletamisel mitmesugustest tüpograafilistest konventsioonidest, nt. sellest, et pärisnimed algavad suurtähega. Samuti teeb oletamise lihtsamaks asjaolu, et sõnastikust puuduvad sõnad kuuluvad teatud väikesesse arvu muuttüüpidesse.

Raskemaks teeb oletamise asjaolu, et pärisnimede käänamisel võib sageli valida, kas käänata sõna eesti keelele omast astmevaheldust kasutades või nime algkuju säilitades. Nt. on juhtunud, et ühe ja sama ajaleheartikli sees kasutatakse nime Fink omastavalise vormina kord astmevahelduslikku vormi Fingi, kord astmevahelduseta vormi Finki. Kui juba inimene ei tea kindlalt, kuidas sõnavorme moodustada, siis on loomulik, et ka automaatne analüüs on raskustes, püüdes omakorda mõistatada, millist vormimoodustamise viisi inimene on kasutanud.
Silpide arvu arvestamise teeb omakorda raskeks asjaolu, et sõna morfoloogilisi omadusi määrava silpide arvu leidmisel tuleb silpe lugeda alates viimasest rõhulisest silbist, sõna rõhku aga ortograafilises tekstis ei märgita. Seetõttu võivad eriti võõrpärisnimede analüüsil ja sünteesil tekkida vead, kuna formaalselt ühesuguse struktuuriga sõnu käänatakse erinevalt, sõltuvalt rõhulise silbi asukohast. Nt. vrd. Vertov (rõhk esimesel silbil) ja Petrov (rõhk teisel silbil): ainsuse osastav on vastavalt Vertovit ja Petrovi. Ehk teisisõnu, kolmesilbiline "ovi"-lõpuline sõnavorm võib tähistada "ov"-lõpulise pärisnime nii omastavat kui osastavat käänet. Viimane on välistatud, kui sõna rõhk on esimesel silbil, kuna seda aga kirjapildist pole näha, siis peab oletaja ta ikkagi välja pakkuma.

6. Analüsaatori vead

Mitte alati ei paku meie programm õiget analüüsivarianti. Järgnevalt kirjeldame, mis liiki vigu võib esineda, et programmi kasutaja oskaks nende suhtes tähelepanelik olla ning neid kas oma töös arvestada või hoopis mingil ad hoc moel ennetada/parandada.
Katsed on näidanud, et sõnastikupõhisel lähenemisel võib õige analüüs puududa analüüsi saanud sõnavormidel (mida on vähemalt 97%) kuni 0,1%-l. Oletamisel, mida rakendatakse ülejäänud 3% sõnavormidele, võib õige analüüs jääda pakkumata aga kuni 10%-l sõnadest. Seega kokku võib kuni 0,1+0,3=0,4%-le sisendteksti sõnadele õige variant puudu jääda.
Sõnastikupõhisel analüüsil on enamlevinud veatüübid järgmised:
1. Sisendtekst pole päris see, mille jaoks analüsaator on mõeldud - puhas tänapäevane kirjakeel, mistõttu sõnad saavad veidra analüüsi. Nt. "puitund" saab analüüsiks "puit_und".
2. Pärisnimi on sarnane mõne üldnimisõna vormiga. Nt. "Rebast" algvormiks pakutakse "Rebane", ehkki tegelikult on algvormiks "Rebas"
Oletamisel tehakse kahte liiki vigu: ei anta sõnale ühtegi õiget analüüsi või antakse õigete hulgas ka valesid analüüse.
Tüüpilisemad vead on järgmised
1. Pakutakse vale sõnaliiki. Nt. suurtäheline sõna määratakse pärisnimeks, ehkki ta ei pruugi seda olla; "budjete" jpt. on samuti määratud nimisõnaks, ehkki nad on hoopis tsitaadid võõr- (antud juhul vene) keelest. Kui sõna on kaks tähte pikk, siis peetakse teda lühendiks. See võib olla ka eksitav, nt. ingliskeelsete eessõnade või hiina nimede puhul.
2. Ei leita õiget algvormi, nt. Loidi puhul ei pakuta algvormiks Loit, vaid Loid.
3. Kuna sõna kuju alusel on raske (kui mitte võimatu) öelda, kus asub sõna rõhk, siis pakub oletaja lisaks õigele mõnikord ka selliseid algvormi kujusid, mis inimesele, kes sõna häädlust teab, paistavad ilmselgelt valed.
Need on probleemid, mille lahendamiseks ei piisa sellest, et üksiksõnade analüüsi täiustada. Perspektiivne oleks hoopis konteksti vaatamine ja selliste sõnade, mille puhul võib kahtlustada vigast analüüsi, muude vormide

otsimine tekstist. Sel juhul saaksime aimu, et "Rebast" algvorm võib olla "Rebas" või et "puitund" peaks olema tegusõna.

7. Ühestaja

Morfoloogiline ühestamine seisneb morfoloogiliselt analüüsitud lause igale sõnale tema võimalike morfoloogiliste märgendite hulgast õige valimises. Näiteks morfoloogiliselt analüüsitud lausest:

```
Mees
   mees+0 //_S_ sg n, //
   mesi+s //_S_ sg in, //
peeti
   peet+0 //_S_ adt, sg p, //
   pida+ti //_V_ ti, //
kinni
   kinni+0 //_D_ //
```

saame peale ühestamist:

```
Mees
   mees+0 //_S_ sg n, //
peeti
   pida+ti //_V_ ti, //
kinni
   kinni+0 //_D_ //
```

Morfoloogilisel ühestamisel lähtutakse järgmisest kahest eeldustest (Merialdo 1994, lk 156):

1. Igale sõnale sobib ainult teatav väike hulk morfoloogilisi märgendeid kõigi võimalike morfoloogiliste märgendite hulgast. See hulk leitakse morfoloogilise analüsaatori abil.
2. Kui sõnal lauses on mitu võimalikku morfoloogilist märgendit, siis lokaalse konteksti põhjal on võimalik määratleda iga sõna jaoks ainus korrektne märgend.

Meie poolt kirjeldatavasse töövahendite komplekti kuuluva ühestaja aluseks on Markovi Varjatud Mudeli (VMM), ingl. k. Hidden Markov Model (HMM) nime all tuntud tõenäosuslik mudel, mille kohta põhjalikumalt vt. (Kaalep, Vaino 1998). Ta tugineb tekstide põhjal tehtud statistikale ega kasuta lingvistile intuitiivselt arusaadavaid reegleid sobivate märgendite valimisel. Me rakendame bigramm-VMMi tema puhtal klassikalisel kujul, mille puhul eeldame järgmist.

1. Lauset ei vaadelda kui sõnade järjestust, vaid kui mingite spetsiaalsete ühestamis-märgendite (M) järjestust. Need on saadud morfoloogiliste märgendite teisendamisel ja neid kasutatakse eelkõige algoritmi paremaks tööks.
2. Kuna sõnal võib olla mitu M, siis konkreetsele lausele võib vastata mitu võimalikku Mide järjestust, aga ainult üks neist on õige.
3. Mõned järjestused on antud keeles tüüpilised, mõned mitte.
4. Võimalikest järjestustest tuleb valida kõige tüüpilisem, s.o. kõige tõenäolisem. See ongi antud lause puhul õige.

5. Uue lause Mide järjestuse tõenäosuse arvutamisel lähtutakse tõenäosustest, mis on leitud varem treenimisfaasis üheselt märgendatud lausete alusel.

Üheste, konteksti sobivate märgendite valimise algoritm on lühidalt järgmine.

1. Teisendame morfoloogilised märgendid ühestamis-märgenditeks M.

2. Arvestame kahte liiki tõenäosusi. Esiteks tõenäosust, et sõnale sobib mingi M, kui me konteksti üldse ei arvesta: nt. tõenäosus, et "veel" on määrsõna, on palju suurem kui see, et ta on sõna "vesi" vorm. Teiseks tõenäosust, et sõnale sobib mingi M, kui talle eelneb mingi konkreetne M: nt. kui sõnale eelneb eessõna, siis tõenäosus, et sõna on nimisõna, on palju suurem kui see, et ta on tegusõna. Ad hoc on kasutusel veel tõenäosuste tabel lause esimeste sõnade jaoks, sest neile ju mingit M ei eelne.

Et leida lause kui M-de järjestuse tõenäosust, tuleb üksikute sõnade M-de tõenäosused liita. Nii saame hulga alternatiivseid M-de järjestusi, millest valime selle, mille tõenäosus on suurim. Vastavad M-d ongi siis need, mis antud juhul sõnadele sobivad. Seega me otsime parimat järjestust, mitte parimat üksiksõna M-i tõenäosust: on võimalik, et parimas järjestuses tuleb mõne sõna puhul valida M, mille tõenäosus polegi maksimaalne.

3. Viimaks teisendame M-d tagasi morfoloogiliste märgendite kujule.

Püüdes minimiseerida statistilisel ühestamisel tehtavaid vigu oleme läinud seda teed, et väga raskete juhtumite puhul loobume ühestamisest sootuks ja jätame mitmesuse alles. Selliseid juhtumed on kokku 13,5% sisendsõnadest. Olulisemad mitmeseks jäetavad sõnagrupid on järgmised

1. nud, tud-lõpulised sõnad, 25% kõigist mitmeseks jäävatest sõnadest. Nende puhul on selle otsustamine, kas tegu on tegusõna või omadussõnaga, lähikonteksti arvestades võimatu

2. Sõna "ta", 16%. Selle puhul jääb otsustamata, kas ta on nimetavas või omastavas käändes.

3. Sõna "on", 13%. Selle puhul jääb otsustamata, kas ta on ainsuses või mitmuses.

4. Sõnad "kui" ja "nagu", 13% kõigist mitmeseks jäävatest sõnadest. Nende puhul jääb otsustamata, kas nad on määr- või sidesõnad.

5. Sõnad "mis" ja "kes", 13%. Nende puhul jääb otsustamata, kas nad on ainsuses või mitmuses.

6. Sõnad, mille algvorm on erinev, aga sõnaliik ja muutevorm ühesugused; nt. "mandri" - manner/mander, "lõi" - looma/lööma; 4%. Sel juhul jääb väljundisse mitu erineva algvormiga varianti. Siin morfoloogiline ühestamine ei saagi aidata, sest probleem on leksikaalne või semantiline.

7. Sõnad "üks" ja "teine", 4%. Nende puhul jääb otsustamata, kas nad on arv- või asesõnad.

8. Muud juhtumid moodustavad 12% kõigist mitmeseks jäävatest sõnadest.

Praegu saab u 3% morfoloogiliselt analüüsitud sõnadest ühestamise tagajärjel vale analüüsi (tüve, sõnaliigi või muu morfoloogilise kategooria osas). Valdav enamus vigu, 1/3, seisneb selles, et nimisõna puhul valitakse homonüümsetest käändevormidest (nimetav, omastav, osastav või lühike sisseütlev) vale variant Kui meid huvitab ainult sõnaliik, siis selle puhul eksitakse 1,7% juhtudest, kui aga ainult algvorm, milles ei eristata osasõnu ega suuri-väikesi tähti, siis selle õige versioon puudub ühestatud tekstis 1,5% juhtudest.

8. Probleemid

Eespool kirjeldasime töövahendeid, mida lingvist saab kasutada. Lingvistika kui humanitaarteaduse omapära on aga see, et tema poolt kasutatavad põhimõisted ja -kategooriad ei ole samal moel täpselt määratletud kui

reaalteadustes. See puudutab ka morfoloogilist analüüsi: nii kasutatavate kategooriate süsteemi kui seda, mis on üldse sõna algvorm.
Eesti keele morfoloogiliseks analüüsiks arvuti abil on praegu kasutusel kaks eri detailsusega kategooriate süsteemi. Üks põhineb Ülle Viksi "Väikesel vormisõnastikul" (Viks 1992) ja tema väikeste modifikatsioonidega

morfoloogiline analüsaator; nimetame teda fs-märgendite süsteemiks. Teine sarnaneb rohkem grammatikatega nagu (Valgma, Remmel 1970) ja (EKG 1995) ning rahvusvahelise standardiseerimisprojekti EAGLES kategooriatega (Monachini, Calzolari 1995) ning teda on kasutatud tekstide käsitsi ühestamisel; nimetame teda kym-süsteemiks.
CG-ühestaja (Puolakainen 1998) ja süntaksi analüsaator (Müürisep 2000) eeldavad, et tekst on märgendatud kym-süsteemis; meie poolt kirjeldatav ühestaja eeldab, et fs-süsteemis.
Kui tekst on märgendatud ühes neist süsteemidest, siis tema teisendamine teise on täisautomaatne. Samas tuleb arvestada, et erinevate süsteemide kasutamine, isegi kui automaatne teisendusprogramm on olemas, tekitab raskusi programmimoodulite sidumisel.
Praktika on mitmete keelte puhul näidanud, et ühestamisel kasutatavate märgendite süsteem on märgendamise täpsuse seisukohalt tähtsamgi kui algoritm või programm ise. Ebasobiva märgendisüsteemi puhul ei oska inimene ega ammugi programm otsustada, kuidas konkreetset sõna tekstis tuleks märgendada. Tulemuseks on ebajärjekindlalt märgendatud tekst, mille kasutaja ei tea, kuivõrd ta seda usaldada võib.
Seega oluliseks probleemiks morfoloogilisel ühestamisel on sobiva märgendussüsteemi valik. See võib tunduda kummaline, sest eesti keele morfoloogia on hästi läbi uuritud. Tegelikult tuleb siiski eristada morfoloogilisi ja süntaktilisi kategooriaid, mida põhimõtteliselt saab eesti keele puhul kasutada, kategooriatest, mida tegelikult on võimalik ühtlaselt ja ühetaoliselt tekstidest eristada. Viimaseid on tunduvalt vähem. Detailselt on vastavaid probleeme käsitletud artiklis (Kaalep jt. 2000), (Puolakainen 2000). Siinkohal piisab tõdemusest, et teoreetilistes käsitlustes nagu (Valgma, Remmel), (EKG) on eesti keele sõnu pahatihti liigitatud sellise detailsusega süntaktilistesse ja semantilistesse klassidesse, mida konkreetses tekstis ka haritud lingvistil ei õnnestu ühtlaselt ja ühetaoliselt määrata. Sellisel juhul on ausam jätta teoreetiliselt võimalik detailne märgendus hoopis tegemata, kui teha seda ebajärjekindlalt.
Omaette probleem morfoloogilise analüüsi ja lemmatiseerimise jaoks on, et erinevad lingvistilised eesmärgid nõuavad erinevaid algvorme. Eesti keele grammatiline traditsioon peab regulaarset tuletust sisaldava sõnavormi algvormiks tuletust sisaldavat vormi, nt. "minemise" algvorm on "minemine". Sõnastike tegemise traditsiooni kohaselt aga regulaarseid tuletisi iseseisvate sõnadena sõnastikesse ei lülitata, seega peab ta algvormiks ilma tuletuseta vormi, nt. "minemise" puhul "minema". Selline algvormi mõiste erinev käsitlus lingvistika kahe haru poolt tekitab probleeme, kui tahame nende poolt kasutatavat keelematerjali, nt. tekstikorpusi ja sõnastikke, omavahel automaatselt ühendada, kasutades selleks morfoloogilist analüsaatorit, mis igale sõnavormile annab ju ainult ühe antud konteksti sobiva analüüsi.

Viited

Ken Church "Unix for Poets". käsikiri
Eesti Keele Grammatika 1995. 1. Toim. M. Erelt; Eesti TA EKI, Tallinn.

Kaalep, H-J. 1997. An Estonian Morphological Analyser and the Impact of a Corpus on its Development. Computers and Humanities. 31: 115-133, 1997.
Kaalep, H-J. 1998. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. Keel ja Kirjandus 1/1998, lk 22-29
Kaalep, H-J., Vaino, T. 1998. Kas vale meetodiga õiged tulemused? Statistkale tuginev eesti keele morfoloogiline ühestamine. Keel ja Kirjandus 1/1998, lk 30-38
Kaalep jt. 2000 Keel ja kirjandus, ilmumas
Koskenniemi, K. 1983. Two-level Morphology: A General Computational Model for Wordform Recognition and Production. Publications of the Dept. Of General Linguistics, University of Helsinki, 11
Merialdo, B. 1994. "Tagging English text with a probabilistic model." Computational Linguistics, 20(2), 155-171.
Monachini M., Calzolari, N. 1995. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and in Corpora and Application to European Languages. EAGLES document EAG-LSG-T4.6/CSG-T3.2, Pisa.
Müürisep, K. 2000 käesolevas kogumikus
Puolakainen, T. 1998. "Eesti keele kitsenduste grammatika morfoloogiline ühestaja." Keel ja Kirjandus, 1, lk. 37-46
Valgma, J., Remmel, N. 1970. Eesti Keele Grammatika. Valgus, Tallinn
Viks, Ü. 1992. Väike vormisõnastik I. Sissejuhatus & grammatika; II. Sõnastik & lisad. Tallinn.
Voutilainen, A., Heikkilä, J., Anttila, A. 1992. Constraint Grammar of English. A Performance-Oriented Introduction. Univ. of Helsinki, Dept. of General Linguistics, No. 21