EESTI KEELE RESSURSSIDE LOOMINE JA KASUTAMINE KEELETEHNOLOOGILISES ARENDUSTÖÖS

HEIKI-JAAN KAALEP

Sisukord

Artiklid **4**
Lühendid **5**
1. Sissejuhatus **6**
2. Taust **8**
3. Keeleressursid **10**
3.1. Korpused **10**
3.1.1 Tartu Ülikooli korpused 11
3.1.2 "1984" 12
3.1.3 Soovitusi tänapäevaste korpuste loomiseks 15
3.2. Sõnastikud **16**
3.2.1 Multext-Easti leksikon 17
3.2.2. ESTMORFi sõnastik 19
4. Teoreetilised küsimused **21**
4.1. Morfoloogiliste kategooriate süsteem **21**
4.2 Lühikese sisseütleva ja vokaalmitmuse kasutamine **22**
4.3 Produktiivsed liitumid eesti keeles **23**
4.3.1 Tuletised 23
4.3.2 Liitsõnad 24
4.4 Sõnajärg **25**
5. Praktilised töövahendid **26**
5.1 ESTMORF, eesti keele morfoloogiline analüsaator **26**
5.2 Ühestaja **27**
6. Kokkuvõte **29**
Abstract **30**
Kirjandus **31**
Elulookirjeldus **34**
Curriculum Vitae **35**

ARTIKLID

I Heiki-Jaan Kaalep. ESTMORF, a Morphological Analyzer for Estonian. Kogumikus H. Õim (toim.) Estonian in the Changing World. Tartu, 1996, lk. 43-98

II Heiki-Jaan Kaalep. An Estonian Morphological Analyser and the Impact of a Corpus on Its Development. Computers and the Humanities 31: lk. 115-133, 1997

III Heiki-Jaan Kaalep. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. Keel ja Kirjandus 1/1998, lk 22-29

IV Heiki-Jaan Kaalep, Tarmo Vaino. Kas vale meetodiga õiged tulemused? Statistkale tuginev eesti keele morfoloogiline ühestamine. Keel ja Kirjandus 1/1998, lk 30-38

V Heiki-Jaan Kaalep, Rene Prillop, Epp Ehasalu. The Role of Internet in Creating, Financing and Integrating Language Resources. Proceedings of the First International Conference on Language Resources and Evaluation. Granada, 1998. Kd. 2, lk 1149-1152

VI Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki-Jaan Kaalep, Vladimir Petkevic, Dan Tufis. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European languages. COLING-ACL '98, Proceedings of the Conference, Kd. 1, lk. 315-319

LÜHENDID

BNS — Baltic News Service
BUUK — BNS uudiste korpus
CES — Corpus Encoding Standard
ESTMORF — eesti keele morfoloogiline analüsaator
MM — morfoloogiline märgend
MVM — Markovi varjatud mudel
TEI — Text Encoding Initiative
TÜKK — Tartu Ülikooli kirjakeele korpus
TÜ — Tartu Ülikool
VVS — Väike vormisõnastik
ÜM — ühestamismärgend

1. SISSEJUHATUS

Väitekirjana esitatud uurimuste aluseks on praktiline töö, mida autor on teinud eesti keele ressursside ja lingvistidele vajalike töövahendite loomisel. Praktilise töö käigus tuli lahendada mitmeid teoreetilisi probleeme, mida väitekirja koosseisus olevad artiklid ka käsitlevad.

Töö kirjeldab eesti keele ressursside, antud juhul korpuste ja leksikonide loomist ning nende kasutamist eesti keele uurimisel, rõhuasetusega praktiliste lingvistiliste töövahendite loomisele.

Laias laastus võibki väitekirjas käsitletava jagada kolme omavahel seotud rühma:

1. Keeleressursid (antud juhul korpused ja sõnastikud) ja nende loomine. Seda teemat käsitlevad suuremal või vähemal määral kõik väitekirja koosseisu lülitatud artiklid.
2. Keeleressursside alusel lingvistiliste töövahendite loomine. Seda käsitlevad (Kaalep 1996), (Kaalep 1997), (Kaalep 1998), (Kaalep, Vaino 1998).
3. Töö käigus esile kerkinud teoreetilised probleemid. Seda käsitlevad (Kaalep 1996), (Kaalep 1997), (Kaalep 1998), (Kaalep, Vaino 1998).

Need teemarühmad on omavahel seotud: keeleressursside põhjal loodi töövahendeid; töövahendid ise aga on samas ka keeleressursside loomise vahendiks; nt. morfoloogiline analüsaator on vahendiks morfoloogiliselt analüüsitud korpuse tegemisel.

Ühelt poolt on korpused selleks aluseks, millele tuginedes töövahendeid teha ja keelt uurida; teiselt poolt suunavad keeleuurimise vajadused ja töövahendite katsetamise vajadus korpuste tegemist: tekstide valikut, märgendamist ja kogumismetoodikat. Ühelt poolt võimaldavad lingvistilised töövahendid keelt paremini uurida; teiselt poolt kerkib nende loomisel üles selliseid lingvistilisi probleeme, mis seni on jäänud teoreetiliste uuringute vaateväljast kõrvale.

Väitekirjas kirjeldatav kajastab tsüklilise protsessi üht (vahe)tulemust, kus korpuste tegemine ja kasutamine, töövahendite loomine ja kasutamine, teoreetiliste probleemide esilekerkimine ja lahendamine toimuvad omavahel seotud astmete kaupa.

Väitekirjas kajastamist leidvad tulemused on saadud Tartu Ülikooli üldkeeleteaduse õppetooli juures töötades ning osaledes mitmetes projektides: Euroopa Komisjoni poolt finantseeritav Copernicus-projekt Multext-East (http://nl.ijs.si/ME/), Avatud Eesti Fondi projektid STYLUS ja KeeleWeb (http://ee.www.ee/).

Kuna tegemist on valdkonnaga, mida iseloomustab kiire areng ja tihe side praktikaga, siis mõned konkreetseid arvutirakendusi puudutavad asjad on praeguseks juba muutunud, võrreldes sellega, kuidas neid käsitletakse artiklites. Need muutused pole aga nii suured ja põhimõttelised, et peaks hakkama artikleid ümber tegema.

Kolmel dissertatsiooni koosseisus oleval artiklil on mitu autorit.

Artikkel (Kaalep, Vaino 1998) kirjeldab eksperimenti, mille tegemiseks oli vaja luua mitmeid teisendus- ja rakendusprogramme. Suure osa neist tegi Tarmo Vaino. Artikkel (Kaalep, Prillop, Ehasalu 1998) puudutab lisaks keeleressursside loomisele ja interneti kaudu kättesaadavaks muutmisele ka programmi Hüperlinker. Viimase loomisel ei ole dissertatsiooni autor kuidagi

osalenud. Artikkel (Dimitrova, Erjavec, Ide, Kaalep, Petkevic, Tufis 1998) kirjeldab kuue keele ressursside loomist. Dissertandi osa piirdub eesti keele ressurssidega.

Dissertatsioon sisaldab ka varem avaldamata tulemusi, millest olulisemad on: soovitused tänapäevaste korpuste loomiseks (osa 3.1.3), eksperiment Multext-Easti leksikoni loomisel ja statistika selle kvaliteedi hindamiseks (osa 3.2.1) ning hinnang teatud grammatiliste vormide kasutatavuse kohta (osa 4.2).

2. TAUST

Väitekirjas esitatav kuulub arvutilingvistikasse; täpsemalt korpuslingvistikasse ja keeletehnoloogiasse.

Arvutilingvistika on keeleteaduse ja informaatika (ehk arvutiteaduse) hübriid, mis tegeleb inimkeele uurimisega nii arvutite abil kui ka arvutite jaoks. Rakenduslik arvutilingvistika keskendub inimkeele modelleerimise praktilistele tulemustele. Selle valdkonna meetodeid, tehnikaid, töövahendeid ja rakendusi nimetatakse sageli ka kokkuvõtliku terminiga "(inim)keeletehnoloogia". Üheksakümnendate aastate algusest on just see rakenduslik pool koos korpuslingvistikaga muutunud järjest olulisemaks.

Korpuslingvistika põhieesmärk on keele uurimine, kasutades suuri koguseid loomulikul viisil esinevaid andmeid (nt. tekste), erinevalt nt. generatiivsest paradigmast, mil piisas uurija isiklikust keeletunnetusest. "Autentse andmestiku" kasutamine iseenesest ei ole uus nähtus lingvistikas, kuid just viimasel ajal on seoses arvutite laiema kasutuselevõtuga saanud võimalikuks uurida keelt ulatuses, millest varem unistadagi ei võinud.

Keeletehnoloogia, nagu teda käsitletakse Euroopa Komisjoni XIII Peadirektoraadi keeletehnoloogia ametlikul koduleheküljel http://www2.echo.lu/langeng/en/lehome.html, on keelealaste teadmiste rakendamine paremate arvutisüsteemide loomiseks:

1. Inimese ja arvuti vahelise suhtluse parandamiseks
2. Informatsiooni paremaks esitamiseks, kasutamiseks, otsimiseks ja analüüsimiseks
3. Inimkeele paremaks mõistmiseks ja töötlemiseks

Keeletehnoloogia annab meile vahendid, et laiendada ja parandada keele kasutusvõimalusi. Ta tugineb seejuures meie teadmistele keelest ja keele funktsioneerimise põhimõtetest, mis on saadud varasema uurimistöö käigus. Uurimistöö tulemusena selguvad nii keeletehnoloogia jaoks lahendamist vajavad probleemid kui ka tehnoloogia, mida kasutades saab keelt mõista ja töödelda.

Praktikas koosneb keeletehnoloogia teatud hulgast võtetest, mis on realiseeritud arvutitarkvarana, ja keeleressurssidest, mis on arvuti abil töödeldav teadmiste kogum. Arvutitarkvara näiteks võib tuua õigekirjakontrolli, terminite otsimise jooksvast tekstist, optilise tekstituvastuse, kõne äratundmise. Keeleressursid on nt. elektroonilised sõnastikud, formaliseeritud grammatikakirjeldused, terminoloogiabaasid ja tekstikorpused. Loomuliku kõne ja keele uurimisega tegelejad on jõudnud arusaamisele, et töökindlate ja tõhusate keeletoodete areng sõltub otsustavalt sellest, kui kättesaadavad on suured adekvaatsed keeleressursid.

Käesoleva dissertatsiooni aluseks olevad artiklid kirjeldavad tööd, mis on tehtud 1991-1998. 1991. aastal olemas olevatest keeleressurssidest olid olulisemad Ülle Viksi "Väikese vormisõnastiku" trükieelne versioon elektroonilisel kujul ja Indrek Heina morfoloogiline analüsaator, mis põhines "Väikesel vormisõnastikul" ja suutis analüüsida lihtsõnu, andes (mitteühese) analüüsi u. 75%-le ajalehe tekstis esinevatele sõnavormidele (Hein 1994). Esimese eesti keele korpuse, miljoni-sõnalise eesti kirjakeele korpuse, loomine algas TÜ eesti keele laboris alles 1991. a. sügisel.

Vajadus uurida eesti keelt uute vahenditega tähendas seda, et need vahendid tuli alles luua. Seejuures tuli luua nii keeleressursid kui arvutitarkvara.

3. KEELERESSURSID

Käesolevas töös käsitletakse kahte liiki keeleressursse: korpusi ja leksikone.

3.1. Korpused

Korpus on keele (teksti või kõne) kogum, mille alusel saab:
1. analüüsida keelt, et tema omadusi kindlaks teha;
2. treenida mingit arvutiprogrammi, et kohandada teda tööks teatud piiritletud olukorras;
3. empiiriliselt kontrollida keele kohta käivat teooriat;
4. testida keeletehnoloogilist võtet või rakendust, et selgitada, kuidas ta töötab praktikas.

On olemas sadadest miljonitest sõnadest koosnevaid rahvuslikke tekstikorpusi, kuid on olemas ka erivajadusteks loodud korpusi. Nt. võib korpus koosneda autojuhtide suulistest vestlustest kõnet mõistva juhtimissüsteemi imitatsiooniga. Sellist korpust kasutatakse selleks, et kindlaks teha kasutaja-poolseid nõudmisi suuliselt juhitavale juhtimissüsteemile.

Korpuste tegemise alases kirjanduses on palju juttu korpuse tegemise põhimõtetest nii tekstide valikul kui nende märgendamisel.

Tekstide valimisel on terve rida aspekte, millele võib tähelepanu pöörata, näiteks: kas valida terviklikud või osatekstid, kas pöörata tähelepanu žanrile (nt. ilukirjandus, õpikud), valdkonnale (nt. ajalugu, geoloogia), tekstide loomise ajale, tekstide levikule (tiraažile); kas valida tekste selle alusel, kui kerge on nende hankimine tehniliselt ja organisatsiooniliselt jne. Ülevaate antud probleemistikust annavad nt. (Muischnek 1998) ja (Hennoste 1996).

Korpuste märgendamisel tuleb otsustada, milliseid tähistusi kasutada, mida märgendada ja mis järjekorras eri asju märgendada.

Sageli on vaja, et elektroonilisel kujul olevad tekstid sisaldaksid eksplitsiitsel kujul veel mingit muud informatsiooni peale selle, mis originaaltekstides esialgselt olemas on.

Märgendust on vaja juba selleks, et oleks üheselt selge, mida mitmesugused trükitehnilised võtted tähendavad. Nt. kaldkiri võib tähistada tsitaati või rõhutamist; taandrida võib tähistada uue lõigu algust või luuletuses uue rea algust; punkt võib tähistada järgarvu, lühendit või lause lõppu. Erinevates trükistes võib sama asja tähistamiseks kasutada erinevaid märke, nt. otsese kõne tähistamiseks võib kasutada erineva kujuga jutumärke või (vanemates tekstides) hoopis mõttekriipsu. Trükitehniliste võtete interpreteerimine märgenduse kaudu puudutab eelkõige teksti struktuuri: jaotust osadeks, pealkirjadeks ja "päris" tekstiks, peatükkideks, lõikudeks, lauseteks, tsitaatideks, loenditeks jms.

Ka juhul, kui tahame tekstile lisada midagi sellist, mida seal varem üldse polnud, nt. anda igale sõnale morfoloogilise analüüsi, märkida tekstis intonatsiooni ja pause, saab seda teha teatud märgendussüsteemi kasutades.

Omaette küsimus on, milline märgendite süsteem valida. Keeletehnoloogias on laialt kasutusel CES (Corpus Encoding Standard) (Ide, Priest-Dorman, Veronis 1996) ja TEI (Text Encoding Initiative) (Sperberberg-McQueen, Burnard 1994). Esineb ka muid, eeskätt üksikute projektide ja/või

institutsioonide spetsiifilisi märgendussüsteeme, kuid nende osatähtsus võrreldes standardsetega on vähenemas, sest viimaste jaoks on olemas järjest kasvav kogus arvutitarkvara, mis võimaldab just standardsete märgenditega tekste mugavalt kasutada ja töödelda.

Sõltumata sellest, mida tahetakse märgendada, on otstarbekas alustada lihtsamast märgendusest, s.t. sellisest, mille lisamine on võimalikult automatiseeritav ja üheselt mõistetav. Alles siis, kui kogu märgendamist vajav tekstide kogum on lihtsama märgenduse saanud, võib alustada keerulisema märgenduse lisamist. Nt. struktuuri märgendamise puhul alustada osadest ja peatükkidest, seejärel märgendada lõigud, loendid ja luuletused (mis võivad olla väliselt kujult päris sarnased) ja alles seejärel laused. Sel moel jagatakse korpuse märgendus tasanditeks. Nt. CESi puhul tasand 1 tähendab, et märgendatud on osad, peatükid ja lõigud, aga mitte laused. Korpus on tervikuna märgendatud just selle tasandini, milleni on märgendatud tema kõige pealiskaudsemalt märgendatud osa.

Eesti keele puhul on kasutatud mitmeid eri viise korpuste kogumiseks (nt. trükitud teksti sisestamine arvutisse käsitsi, flopi-ketastega tekstide toomine, internetist kopeerimine, e-posti kaudu tekstide saamine) ja märgendamiseks (nt. TEI ja CES eri detailsusega, märgendamata, oma unikaalne märgendussüsteem).

3.1.1 Tartu Ülikooli korpused

Tartu Ülikoolis on tehtud ja tegemisel mitmeid erinevaid korpusi, millest dissertatsiooni autor on koordineerinud kahe tegemist: alates 1995. a. Tartu Ülikooli kirjakeele korpuse (TÜKK) tegemist ja alates 1994. a. BNSi uudiste korpuse (BUUK) tegemist. Neid korpusi iseloomustab suhteliselt suur maht (vähemalt miljon sõna) ja suhteliselt pealiskaudne märgendus.

TÜKKi loomine algas 1991. a. sügisel TÜ eesti keele laboris (Hennoste 1996). 1991-1994 sisestati arvutisse kogu ajakirjanduse osa ja osa ilukirjandust ning märgendati nad, kasutades TEI-sarnast ebastandardset märgendust. 1995-1996 sisestati ülejäänud korpus, märgendati ta TEI järgi ära kuni lause tasandini ning tehti vabalt kättesaadavaks interneti kaudu (http://www.cl.ut.ee/). Suure töö tegid seejuures ära Mare Koit, Riina Mosna, Kadri Muischnek, Heili Orav, Leho Paldre, Urve Talvik, Tarmo Vaino ja Viire Villandi. Alates 1997. a. on võimalik kasutada lisaks TEI järgi märgendatud ja nn. puhta teksti versioonile ka morfoloogiliselt märgendatud versiooni. Viimase tegi Leho Paldre, kasutades morfoloogilist analüsaatorit ESTMORF.

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 on selgunud, et paljudel juhtudel ei ole käsitsi tehtud märgendusest kasu. Nt. morfoloogilise analüsaatori loomisel oli parem kasutada 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. BUUK on kasutatav ainult uurimisotstarbel, ainult TÜ üldkeeleteaduse õppetoolis.

3.1.2 "1984"

G. Orwelli "1984" käsitleme eraldi, sest tegu on suhteliselt väikesemahulise, kuid põhjalikult märgendatud korpusega, mille paralleelne versioon eksisteerib bulgaaria, inglise, rumeenia, sloveeni, tšehhi ja ungari keele jaoks.

"1984" on kasutatav CD pealt (Erjavec, T., Lawson, A., Romary, L. 1998), (http://nl.ijs.si/ME/). Tema tehnilist ülesehitust kirjeldavad (Ide 1996), (Erjavec 1997) ja (Priest-Dorman, Erjavec, Ide, Petkevic 1997). "1984" korpuse loomisel on kasutatud mitmeid keeletehnoloogilisi vahendeid: eesti keele morfoloogilist analüsaatorit ESTMORF (Kaalep 1998), ühestajat (Puolakainen 1998), lausestajat ja joondajat (http://www.issco.unige.ch/) ning mitmeid spetsiaalselt antud korpuse märgendamiseks loodud programme. Kogu märgendus on ka käsitsi üle kontrollitud.

Suure töö eestikeelse "1984" korpuse loomisel on teinud Greg Priest-Dorman Vassari kolledžist (USA) ja Kadri Muischnek, Heili Orav, Leho Paldre, Viire Villandi jt. TÜ üldkeeleteaduse õppetooli töötajad.

"1984" on omakorda kasutatud keeletehnoloogilises arendustöös: just tema alusel on treenitud eesti keele statistilist ühestajat.

Kuna G. Orwelli "1984" elektroonilist versiooni ei õnnestunud 1995. aastal leida, siis on ta raamatu põhjal uuesti sisestatud. "1984" sisaldab 80 000 sõna; ta koosneb kolmest osast ja ühest lisast. Osad on omakorda jaotatud peatükkideks.

"1984" on kolmes elektroonilises versioonis: nn. normaalversioonina, paralleelkorpusena ja morfoloogiliselt analüüsituna ning ühestatuna.

1. Normaalversioon

Vaatame üht lõiku originaalist:

Tõeministeerium - uuskeeles Tõmin - erines rabavalt kõigest muust, mida oli näha. See oli tohutu kiiskavvalgest betoonist püramiidne ehitis, mis kerkis astanguliselt 300 meetri kõrgusele. Sealt, kus Winston seisis, seletas silm veel parajasti valgel seinal elegantses kirjas ilutsevat Partei kolme loosungit:*
SÕDA ON RAHU
VABADUS ON ORJUS
TEADMATUS ON JÕUD

Normaalversioon, mis antud lõigust on märgendatud CES-i kohaselt (Ide, Priest-Dorman, Véronis 1996) , on selline:

```
<p id="Oet.1.2.7">
<s id="Oet.1.2.7.1">
<name type=org>
T&otilde;eministeerium
</name>
 &mdash;
```

<name type=language>
uuskeeles
</name>
<ptr target=oet.N1 rend=asterisk>
<name type=org>
Tõmin
</name>
— erines rabavalt kõigest muust, mida oli näha.
</s>
<s id="Oet.1.2.7.2">
See oli tohutu kiiskavvalgest betoonist püramiidne ehitis, mis kerkis astanguliselt
<num>
300
</num>
meetri kõrgusele.
</s>
<s id="Oet.1.2.7.3">
Sealt, kus
<name type=person>
Winston
</name>
seisis, seletas silm veel parajasti valgel seinal elegantses kirjas ilutsevat
<name type=org>
Partei
</name>
kolme loosungit:
<q id="Oet.1.2.7.3.3" rend=CA type=slogan>
Sõda on rahu
</q>
<q id="Oet.1.2.7.3.4" rend=CA type=slogan>
Vabadus on orjus
</q>
<q id="Oet.1.2.7.3.5" rend=CA type=slogan>
Teadmatus on jõud
</q>
</s>
</p>

Normaalmärgendus on tehtud kuni lausete tasandini, s.t. et raamatu struktuuri osas on märgendatud osad, pealkirjad, peatükid, lõigud, laused, tsitaadid, luuletused, loendid, esiletõstetud tekst (nt. kaldkiri) ja joonealused märkused. Kõik struktuurselt märgendatud osad on varustatud identifikaatoritega, et oleks võimalik eri keelte tekste omavahel siduda. Algul märgendati veel käsitsi lühendeid, kuupäevi, nimesid, numbreid, tiitleid, muukeelseid sõnu ja otsest kõnet. Hiljem sellest loobuti, sest töömaht osutus liiga suureks ja sellise märgenduse vajalikkus on kaheldav. Tulemuseks on see,

et mittestruktuurne märgendus on tehtud ainult esimese osa esimeses peatükis ja sealgi mittejärjekindlalt.

2. Paralleelkorpus

Erinevalt mõnest teisest paralleeltekstist nagu SCLOMB (Yli-Vakkuri 1993), kus on püütud tõlketeksti faili otse siduda originaaliga (nt. kirjutades tõlketeksti lausete juurde originaali lausete numbrid), sisaldab "1984" paralleelkorpusena ainult viitasid osade, lõikude, lausete ja loendite identifikaatoritele.

Näiteks eestikeelse "1984" esimese osa esimese peatüki neljanda lõigu laused 3 ja 4 on originaalis vastavalt laused 3 ja 4 ning 5:

```
<link xtargets="Oet.1.2.4.3 ; Oen.1.1.4.3 Oen.1.1.4.4">
<link xtargets="Oet.1.2.4.4 ; Oen.1.1.4.5">
```

Pärast viitade abil lausete leidmist saame paralleelteksti:

<Oet.1.2.4.3>Mustavuntsiline nägu vahtis vastu iga nurga pealt, ka vastasmaja fassaadilt.
<Oen.1.1.4.3>The blackmoustachio'd face gazed down from every commanding corner.<Oen.1.1.4.4>There was one on the house-front immediately opposite.

<Oet.1.2.4.4>" Suur Vend valvab sind," ütles kiri, ja tumedad silmad vaatasid sügavalt Winstonile silma.
<Oen.1.1.4.5>" Big Brother is watching you," the caption said, while the dark eyes looked deep into Winston's own.

3. Morfoloogiliselt märgendatud ja ühestatud variant.

Vaatame osalauset *Oli külm selge aprillipäev,*. See on morfoloogiliselt märgendatult ja ühestatult järgmine:

```
<orth>Oli</orth>
<disamb><base>olema</base><msd>Vmii3s-an</msd>
<ctag>VM3</ctag></disamb>
<lex><base>olema</base><msd>Vmii3s-an</msd></lex>
<lex><base>olema</base><msd>Vaii3s-an</msd></lex>
</tok>
<tok type=WORD>
<orth>k&uuml;lm</orth>
<disamb><base>=</base><msd>A-p-sn</msd>
<ctag>ASN</ctag></disamb>
<lex><base>=</base><msd>A-p-sn</msd></lex>
<lex><base>=</base><msd>Nc-sn</msd></lex>
</tok>
<tok type=WORD>
<orth>selge</orth>
```

<*disamb*><*base*>=</*base*><*msd*>A-p-sn</*msd*>
<*ctag*>ASN</*ctag*></*disamb*>
<*lex*><*base*>=</*base*><*msd*>A-p-sg</*msd*></*lex*>
<*lex*><*base*>=</*base*><*msd*>A-p-sn</*msd*></*lex*>
</*tok*>
<*tok type=WORD*>
<*orth*>aprillipäev</*orth*>
<*disamb*><*base*>=</*base*><*msd*>Nc-sn</*msd*>
<*ctag*>NCSN</*ctag*></*disamb*>
<*lex*><*base*>=</*base*><*msd*>Nc-sn</*msd*></*lex*>
</*tok*>
<*tok type=PUNCT*>
<*orth*>,</*orth*>
<*ctag*>COMMA</*ctag*>
</*tok*>

Iga sõna puhul on esitatud:

- tekstis esinev sõnavorm (märgend <*orth*>),
- morfoloogilise analüüsi tulemused (märgend <*lex*>), milles on omakorda algvorm (märgend <*base*>), mis juhul, kui algvormi kuju on sama mis sõnavormil, on esitatud '=' kujul, ja morfo-süntaktiline kirjeldus (märgend <*msd*>),
- ühestamise tulemus (märgend <*disamb*>), mis on üks morfoloogilise analüüsi tulemustest (märgendid <*base*> ja <*msd*>), ja ühestamismärgend (märgend <*ctag*>)

Kirjavahemärkide puhul on näidatud, et tegu on kirjavahemärgiga ning neil morfoloogilist analüüsi pole; on ainult ühestamismärgend.

3.1.3 Soovitusi tänapäevaste korpuste loomiseks

Praeguseks on juba selgunud mõned aspektid, millele tuleks tähelepanu pöörata korpuste kui keeletehnoloogia jaoks vajaliku materjali kogumisel.

1. Tekstide valik peaks olema orienteeritud võimalikult tänapäevase keele kajastamisele. TÜKKi kasutusvõimalusi vähendab oluliselt see, et ta sisaldab vananenud keelt. Mitmed TÜKKi osad tooksid keeletehnoloogilistele algoritmidele isegi kahju, kui neid kasutada sõnavara ja/või süntaksi allikana, nt. eriti propaganda ja suur osa ajakirjandusest. Seega võiks öelda, et ajakohasuse mõttes on hea *on-line* kogumine; samuti selle kogumine, mis on saadaval internetis.
2. Märgendus peaks olema kogu korpuse ulatuses ühtlane. Nt ei tohiks lausete märgendamisel panna otsese kõne lauset ja saatelauset kord kaheks omaette lauseks, kord üheks (nagu nt. TÜKKi ajakirjanduse-osas on tehtud).
3. Märgendamisel tuleb alustada lihtsamast struktuursest märgendusest (osad, peatükid, lõigud) ja mitte liikuda keerulisemale märgendusele enne, kui kogu ettevõetud korpus on lihtsamal tasemel märgendatud. Negatiivse näitena võib siin tuua selle, et TÜKKis märgendati käsitsi lühendeid, pärisnimesid ja numbreid. Nende märgendamine kuulub

tegelikult morfoloogilise märgendamise etappi ja on sellisena automaatselt tehtav. Käsitsi märgendamine tähendab seda, et märgendusse tekib juhuslikke vigu ning hiljem tuleb vaeva näha spetsiaalsete filtritega, mis morfoloogilise märgendamise etapil varem käsitsi märgendatud sõnad vahele jätaks. Lihtsam on teha nii, et algul märgendada automaatselt mingi tasand ära (nt. anda sõnadele morfoloogiline analüüs) ja seejärel (pool)automaatselt need osad, mis automaatsest tööst kõrvale jäid, nt. haruldased pärisnimed, lühendid, trükivead.

4. Vajalikud on mitmesugused erinevad tekstid: nii toimetajate käest läbikäinud kui toimetamata (nt. tüüpiliste trükivigade leidmiseks)
5. Spetsiaalselt tuleb tähelepanu pöörata sellele, et korpus oleks kättesaadav laiemale publikule, nt. CD või interneti kaudu. Võimalike autoriõiguse alaste takistuste ettenägemine mõjutab tekstide valikut. Korpuse tegemine kättesaadavaks väljaspool kitsast tegijate ringi nõuab omakorda lisapingutusi formaalse ühtluse ning dokumentatsiooni loomisel, mida tuleks korpuse tegemisel algusest peale arvestada.
6. Vajalik on korpuse pidev täiendamine ja leitud vigade parandamine. Korpuse loomine ei ole päris ideaalselt etappideks jagatav (vastuolu nõudega nr. 3). Igal etapil võib välja tulla eelmise etapi vigu: trükivigu, märgenduse vigu, mõne programmi töö vigaseid tulemusi.
7. Kuna märgendamine on väga töömahukas tegevus, siis on mõtet märgendada nii vähe kui võimalik, ehkki nii palju kui vajalik. Märgendus, ilma milleta on raske korpusi kasutada, hõlmab tekstide allikaid (mis kasu on andmetest, kui ei ole teada, kust nad pärinevad ja kui usaldusväärsed nad on?) ja struktuuri (peatükid, lõigud, laused, sõnad).

3.2. Sõnastikud

Elektrooniline sõnastik e. *leksikon* on sõnade ja nende kohta käivate teadmiste kogum. Need teadmised võivad olla nt. morfoloogia, fonoloogia, tähenduse kohta. On raske leida keeletehnoloogilist rakendust, milles üldse ei kasutata mingit leksikoni. Elektroonilised sõnastikud erinevad traditsioonilistest, inimese jaoks mõeldud (paber)sõnastikest nii oma struktuuri kui sisu poolest, mistõttu elektrooniliste sõnastike tegemine traditsiooniliste alusel või lausa nullist on oluline osa keeletehnoloogilisest arendustööst.

Leksikonidest käsitletakse antud töös kahte morfoloogiliseks analüüsiks mõeldud leksikoni:

1. korpuse baasil loodud leksikoni.
2. eesti keele morfoloogilise analüsaatori aluseks olevat leksikoni

Milleks on meil vaja mitut samaotstarbelist leksikoni? Põhjused on eelkõige tehnoloogias: üks leksikon on tehtud lihtsa struktuuriga, et lihtsad programmid saaksid teda kasutada, teine aga spetsiaalselt kohandatud kvaliteetseks morfoloogiliseks analüüsiks programmi ESTMORF poolt.

3.2.1 Multext-Easti leksikon

Paljud keeletehnoloogilised rakendused vajavad sõnavormide analüüsimise vahendeid. Nt mõnikord on vaja abstraheeruda inflektsioonilistest variantidest, nii et nt. *minna, lähen, läksin* käsitletakse kõiki sõna *minema* variantidena. Mõnikord on aga soovitav kasutada informatsiooni, mida puhtas tekstis ei leidu, nt. et *lähen* on kindla kõneviisi oleviku ainsuse esimene pööre sõnast *minema*.

Esimest ülesannet nimetatakse lemmatiseerimiseks, teist morfoloogiliseks analüüsiks; ja mõlemaid saab lahendada, kasutades spetsiaalseid programme. (Keele)tehnoloogiline küsimus, mis seejuures esile kerkib, on see, et meid ei rahulda tegelikult lihtsalt mingi programmi olemasolu, vaid programmi olemasolu konkreetse riist- ja tarkvaraplatvormi jaoks. Sellest seisukohast vaadates peaks morfoloogilise analüüsi ja lemmatiseerimise programm olema võimalikult lihtne ja universaalne, et ta oleks kergesti muudetav, kohendatav ja sobiks eri keeltele. Ainus viis seda saavutada on eristada analüüsi algoritm ja andmed; algoritm omakorda peaks olema keelest sõltumatu ja kergesti ümberprogrammeeritav, andmed keelele omased ja kergesti kasutatavad. Praktikas tähendab see seda, et tuleb kasutada lihtsa struktuuriga leksikaalset andmebaasi — sõnastikku — ja analüüsi asemel võimalikult lihtsat sõnastikust otsimist.

Mitmete Lääne- ja Ida-Euroopa keelte jaoks piisab kolme-veerulisest tabelist — leksikaalsest andmebaasist (vt. tabel 1), mille veergudes on kirjas:

1. sõnavormid,
2. neile vastavad algvormid,
3. grammatiline info, mida konkreetne sõnavorm esindab.

sõnavorm	algvorm	morfo-süntaktiline kirjeldus
lähen	minema	verb, kindel kõneviis, olevik, ainsus, 1 pööre

Tabel 1. Lihtne leksikaalne andmebaas

Kui tahame leida mõne sõnavormi algvormi ja/või grammatilist infot, siis tuleb sellisest sõnastikust otsida üles sõnavorm (see on esimeses veerus). Sama kirje teises veerus ongi siis algvorm ja kolmandas grammatiline info.

Selline leksikon peab sisaldama piisavalt palju sõnavorme, et katta jooksvas tekstis esinevaid sõnu. Samal ajal peab leksikon olema nii väike, et ta on arvutustehnika praeguste võimaluste juures ikka kasutatav.

Kas eesti keele jaoks õnnestub kasutada sellist ülilihtsat morfoloogilise analüüsi meetodit, arvestades eesti keele morfoloogilist keerukust?

Esimene võimalus luua sõnavormide leksikon oleks kasutada mingit olemasolevat sõnatikku ja genereerida sellest kõikvõimalikud sõnavormid. Kui me kasutaksime Väikest vormisõnastikku (Viks 1992), siis saaksime 35 000 algvormist genereerida 1,2 miljonit sõnavormi. Tuletiste ja liitsõnade lisamine viiks sõnade hulga miljarditesse. Seega lihtsalt kõikvõimalike sõnavormide genereerimine ei oleks praktiliselt mõistlik.

Teine võimalus oleks võtta aluseks mingi hulk eestikeelseid tekste, teha nende alusel sõnavormide leksikon ja loota, et saadud sõnavormide hulk katab küllalt suure osa ka tundmatute tekstide sõnavormidest. Allpool kirjeldatakse ühte sellist katset.

Leksikoni tegemiseks võeti 150 000 sõna ulatuses tekste TÜKKi ilukirjanduse, ajakirjanduse ja teaduse osast, kokku 450 000 sõna ulatuses. Nad analüüsiti ESTMORFiga ära ja saadi sõnastik, milles oli 118 000 erinevat kirjet. 1. veerus oli 81 000 erinevat sõnavormi ja 2. veerus 43 000 erinevat algvormi. Kirjete ja sõnavormide arvu erinevus tuleneb sellest, et paljudel sõnavormidel on mitu võimalikku algvormi ja/või grammatilist tõlgendust (vt. tabel 2).

aega	aega	kaassõna
aega	aeg	nimisõna, ainsus, osastav
aega	aeg	nimisõna, ainsus, lühike sisseütlev

Tabel 2. Sõnavormi *aega* kirjed lihtsas leksikaalses andmebaasis

Selline leksikon ei sisalda kõigi sõnade täisparadigmasid, küll on ta väga sobiv nendesamade tekstide analüüsimiseks, mille põhjal ta on koostatud. Et hinnata tema sobivust ka tundmatute tekstide analüüsiks, tehti katse G. Orwelli "1984ga" . Tulemused on tabelis 3.

	tekstis sõnu	erinevaid sõnavorme
kokku	80 000	17 900
neist sisaldus leksikonis	68 600	10 400
sama, protsentuaalselt	86	58
neist puudus leksikonist	11400	7500
sama, protsentuaalselt	14	42

Tabel 3. "1984" analüüs sõnavormide leksikoni abil

Nagu näha, katab loodud sõnavormide leksikon tundmatut teksti paremini kui morfoloogiline analüsaator, mis tunneb ära ainult lihtsõnade kõikvõimalikud vormid (Hein 1994): viimane tundis ära 75% jooksva teksti sõnadest. Praktilistel eesmärkidel on morfoloogiline analüsaator, mille aluseks on selline leksikon ja ainsaks meetodiks sealt sõnavormi otsimine, oma katvuse poolest siiski kasutuskõlbmatu.

Esimene pähetulev idee saadud leksikoni parandamiseks on järgmine: tuleks laiendada saadud leksikoni sel moel, et genereerida kõigist lemmadest koguparadigmad. Sel juhul saaksime leksikoni, mis katab tundmatut teksti kindlasti paremini. Küsimuseks on, kui palju paremini: kui palju tundmatuid sõnavorme on leksikonis olevate lemmade tundmatud vormid ja kui palju on mingite uute lemmade, s.h. tuletiste ja liitsõnade, vormid?

Et seda kindlaks teha, tehti järgmist. Analüüsiti "1984" morfoloogiliselt, kasutades ESTMORFi. Tulemused on tabelis 4.

	tekstis sõnu	erinevaid sõnavorme
kokku	80 000	17 900
analüüsitud	78 200	17 500
sama, protsentuaalselt	98	98
tundmatud	1800	400

sama, protsentuaalselt	2	2

Tabel 4. "1984" analüüs ESTMORFi abil

Seejärel eraldati sõnad, mis lihtsa sõnavormide leksikoni puhul jäid tundmatuks, ESTMORFi poolt aga ära analüüsiti. Leksikonis puudunud sõnavormide jagunemist oma päritolu poolest kirjeldab tabel 5.

tekstis sõnu kokku	tekstis sõnu		erinevaid sõnavorme	erinevaid sõnavorme kokku
9600	5500	olemasolevate lemmade uued vormid	4100	7100
	4100	uute lemmade vormid	3000	

Tabel 5. Sõnavormide leksikonis puudunud sõnade moodustusviis

Näeme, et olemasoleva leksikoni laiendamine nii, et genereerime kõigi seal olevate lemmade paradigmad, annaks meile "1984" puhul katvuseks ligi 95%. Tundmatuks jääb 2,5 korda rohkem sõnu kui ESTMORFi puhul, kuid katvus ulatub siiski tasemele, mida võib pidada aktsepteeritavaks (Vuotilainen, Heikkilä, Anttila 1992). Teiste sõnadega, eesti keele puhul oleks võimalik luua antud leksikoni põhjal küllalt hea morfoloogiline analüsaator, kui me võtaksime aluseks olemasolevad lemmad ja piirduksime sõnamuutusega, tegemata katsetki analüüsida sõnastikust puuduvaid tuletisi ja liitsõnu.

Leksikoni laiendamist kõigi võimalike sõnavormidega pole siiski tehtud, sest tulemuseks oleks praeguste arvutiressursside jaoks liiga suur tabel. Selle asemel on tehtud järgmist.

Et sõnavormide leksikon oleks kooskõlas korpusega, mida me põhjalikult analüüsime ja märgendame — G. Orwelli "1984ga", siis on sinna lisatud kõik uued sõnavormid, mida ESTMORF suutis analüüsida. Välja jäid Uuskeele sõnad (nt. *prole*) ja briti pärisnimed (nt. *Syme*). Tulemuseks on leksikon, milles on 130 500 kirjet, 89 000 erinevat sõnavormi ja 45 000 erinevat algvormi. Leksikon on CDl (Erjavec, Lawson, Romary 1998).

3.2.2. ESTMORFi sõnastik

Morfoloogiline analüsaator peab olema võimalikult täpne. See tähendab, et ta peaks võimaldama analüüsida piisavat hulka reaalsetes tekstides esinevaid sõnu, kuid samas ei tohiks ta analüüsida selliseid sõnu, mida tekstides ei esine (ja mis selles mõttes antud keelde ei kuulugi), nagu nt. käibelt kadunud sõnad või mitte juurdunud uudissõnad. Morfoloogilise analüsaatori täpsust mõjutab nende sõnade valik, mis tema sõnastikku kuuluvad. Keelele mitteomaste sõnade olemasolu sõnastikus toob kaasa riski, et kirjavigadega sõnad, lühendid ja muude keelte sõnad (nt. tsitaatides) analüüsitakse valesti kui antud keele normaalsed sõnad.

Kuldne kesktee sõnastiku katvuse ja täpsuse vahel on saavutatav ainult sel teel, et me kontrollime sõnastikku (ja morfoloogilist analüsaatorit) reaalset keelekasutust esindavate tekstide peal. Praeguseks on ESTMORFi sõnastik järgmine.

ESTMORFi sõnastikus on 38 000 sõna. Ta põhineb Väikese vormisõnastiku (VVS) elektroonilisel versioonil (Viks 1992), milles on 35000 lihtsõna. Võrreldes ESTMORFi sõnastikku VVSiga näeme, et sinna on lisatud:

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 neist tuletatud genitiivatribuuti, s.h. u. 70 võõrpärisnime, mis koosnevad mitmest sõnast nagu *New York*.
4. Ligikaudu 300 lühendit.

VVSist on eemaldatud:

1. Ligikaudu 1800 vananenud või murdesõna
2. Ligikaudu 2700 liigset tuletist (VVS sisaldab palju produktiivseid tuletisi)

4. TEOREETILISED KÜSIMUSED

4.1. Morfoloogiliste kategooriate süsteem

Üks aspekt, mis nõuab keeletehnoloogias ja arvutilingvistikas erilist tähelepanu, on kasutatavate kategooriate formaliseeritus, üheselt mõistetavus, täielikkus ja mittevastuolulisus. Antud töö raames tuli kokku puutuda morfoloogiliste kategooriate süsteemiga, mida kasutatakse sõnavormide morfoloogilisel klassifitseerimisel.

Eri allikad annavad eesti keele morfoloogiliste kategooriate ja selle kohta, kuidas sõnad nende vahel jagunevad, eri pildi. Eriarvamused hõlmavad esiteks seda, kas eesti keele sõnad jagunevad sõnamuutumise seisukohalt 3 rühma — käändsõnad, pöördsõnad ja muutumatud sõnad (Viks 1992); või tuleb eri rühmadena käsitleda veidi teisel moel muutuvaid sõnu — võrdlussõnu nagu *ruttu — rutem* või kohakäändesõnu nagu *peale — peal — pealt* (Eesti Keele Grammatika 1995). Teiseks on eri autorid eri arvamusel selle suhtes, millistesse sõnaliikidesse võivad sõnad jaguneda. Muutumatute sõnade puhul on (Valgma, Remmel 1970), (Viks 1992) ja (Eesti Keele Grammatika 1995) seisukohtadel, mida kirjeldab tabel 6.

Näide	Valgma, Remmel 1970	Viks 1992	EKG 1995
eri	omadussõna	omadussõna	omadussõna
balti	omadussõna	(genitiiv)atribuut	omadussõna
sees	määrsõna	määrsõna	määrsõna
siin	määrsõna	määrsõna	asesõna
vist	määrsõna	määrsõna	rõhumäärsõna
tagasi	määrsõna	määrsõna	abimäärsõna
plehku	ei käsitle	sõnaliik puudub	ei käsitle

Tabel 6. Sõnade klassifitseerimine eri allikates.

Kolmandaks ei ole eri autorid üksmeelel selles osas, milliseid käänd- ja pöördsõnavorme eristada. Nt. (Viks 1992) eristab omaette käändena lühikest sisseütlevat e. aditiivi, mida muud autorid ei tee (Eesti Keele Grammatika 1995). Pöördsõnade puhul eristab (Eesti Keele Grammatika 1995) möönvat kõneviisi, mida mitmed muud autorid, nt. (Valgma, Remmel 1970) ja (Viks 1992) ei tee; (Viks 1992) aga kaudse kõneviisi mineviku umbisikulist vormi (nt. *elama — elatuvat*); (Eesti Keele Grammatika 1995) lubab ainult vormi *olevat elatud*.

Erinevad morfoloogiliste kategooriate süsteemid on tingitud sellest, et uurijad keskenduvad erinevate aspektide kirjeldamisele. Oleks hea, kui eksisteeriks üks selline kategooriate süsteem, millest kõik ülejäänud on tuletatavad. Paraku praegu sellist pole ja ei tea, kas seda saabki teha. Seetõttu on ka avalikult kasutatavates eesti keele korpustes ning lingvistilistes töövahendites praegu kasutusel kaks erinevat morfoloogiliste kategooriate süsteemi: TÜKKis kasutatakse väikeste modifikatsioonidega Väikese vormisõnastiku süsteemi

(http://www.filosoft.ee/html_morf_et/morfoutinfo.html), "1984" puhul aga Multext-Easti süsteemi (Erjavec, Monachini 1998).

Esimene sobib paremini spetsiifiliselt sõnamuutuse käsitlemiseks, teine vastab oma kirjeldusmehhanismi poolest aga rahvusvahelisele standardile EAGLES (Monachini, Calzolari 1995), (Bel, Calzolari, Monachini 1995), mis võimaldab kasutada eesti keele peal mitmeid muudes maades loodud lingvistilisi töövahendeid ning testida nende sobivust eesti keelele.

Eraldi tuleks vast mainida sedagi, et lisaks traditsioonilistele lingvistilistele kategooriatele tuleb keeletehnoloogias anda mingi analüüs ka tekstis esinevatele mittesõnadele: valemitele, kirjavahemärkidele, lühenditele. Nii VVSil põhinev süsteem kui ka Multext-Easti süsteem selliseid kategooriaid ka sisaldavad.

4.2 Lühikese sisseütleva ja vokaalmitmuse kasutamine

Mõningaid grammatilisi vorme peavad eesti keele kasutajad keeleomasteks, mõningate kasutamist aga väldivad. Teoreetiline küsimus kasutatavatest vormidest saab kõige täpsema ja ammendavama vastuse sõnastikus, mis teatud vormide kasutamist aktsepteerib, teatud vormide oma aga mitte.

Nii nagu morfoloogilise analüsaatori sõnastik ei tohi sisaldada sõnu, mis on nii haruldased ja antud keele kasutajale veidrad, et neid peetakse trükivigadeks või võõra keele sõnadeks, peaks olema ettevaatlik ka võimalike sõnavormide hulga lubamisel. Korpuste põhjal tehtud statistika alusel on ESTMORFi leksikoni kohandatud eesti keele reaalse kasutusega, mis puudutab lühikese sisseütleva vorme ja vokaalmitmust.

VVS lubab väga paljudele käändsõnadele lühikest sisseütlevat käänet, nt. *kehasse* e. *kehha*. ESTMORFi sõnastikust on eemaldatud selliseid reaalsele keelekasutusele ilmselt mitte vastavaid vorme ümmarguselt 8000 sõna puhul. Valdavalt eemaldati vormid, mille puhul VVSis oli märgitud, et nende moodustamine on ebatõenäoline, v.a. *sepp*-tüüpi sõnade lühike sisseütlev, mis langeb kokku ainsuse osastavaga. Selline eemaldamine muude tüüpide puhul ei olnud aga automaatne; kõik sõnad vaadati eraldi üle ja mõnekümnel juhul otsustati VVSis märgitud ebatõenäolised vormid (nt. *torru*) siiski alles jätta.

VVS lubab ka vokaalmitmust väga paljudele sõnadele, nt. *atradel* e. *adrul*. ESTMORF on VVSist rohkem kui 7000 sõna puhul, millest valdav enamus kuulub *sepp*-tüüpi, selles osas rangem.

Omaette küsimuseks on, kas käändsõnade vokaalmitmust võib esineda nelja viimase käände (rajav, olev, ilmaütlev, kaasaütlev) puhul. (Viks 1992) ja (Peebo 1997) arvavad, et reeglina mitte. Samas võib TÜKKist leida sõnu nagu *põlvini, silmini, õnnelikena, surmväsinuina, viljelejaina, soosikuina, võrdväärseina*. Reegliks näib olevat, et *õnnelik*-tüüpi sõnadel on vokaalmitmus võimalik kõigi nelja viimase käände puhul; i-mitmusega sõnadel on vokaalmitmus võimatu kahe viimase käände puhul, tüvemitmusega sõnadel aga tõepoolest nelja viimase käände puhul. Erandeiks on üksikud sõnad nagu *silmini, põlvini, kõrvuni, õluni, pilvini, rinnuni, ladvuni*, millel pole vokaalmitmust kolme viimase käände puhul.

4.3 Produktiivsed liitumid eesti keeles

Ilma korpuste ja automaatse morfoloogilise analüüsi võimaluseta on raske vastata küsimustele:

1. Kui produktiivne on sõnamoodustus reaalsetes tekstides?
2. Millised on tuletuse ja liitsõnade moodustamise mallid ja millised neist on produktiivsed?

Varasemad uurimused nagu (Kask 1967), (Kull 1967), (Kasik 1984) ja (Kasik 1992) annavad kasulikke vihjeid mõlemale küsimusele vastamiseks, aga ei ole morfoloogilise analüsaatori loomisel koheselt kasutatavad. Lisaraskusi tekitab asjaolu, et sõnamoodustust kirjeldatakse kui sünteesiprotsessi, meid huvitab aga analüüs. Eraldi probleemiks on veel see, et liitsõnade moodustust kirjeldatakse kui kahe komponendi liitmist, samas kui reaalsetes tekstides esineb kuni 5-komponendilisi liitsõnu. Ei ole selge, kuivõrd võib keerulisema struktuuriga sõnade puhul kasutada rekursiivselt samu reegleid, mida kasutatakse 2-komponendiliste sõnade puhul.

4.3.1 Tuletised

Umbes 8% kõigist sõnedest eestikeelses tekstis on tuletised; ajakirjandustekstis 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.

4.3.2 Liitsõnad

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 selline, nagu tabelis 7.

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%

Tabel 7. Liitsõna-mallide sagedasemad tüübid

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 komponentide 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 on jõutud järgmise variantide proovimise järjekorrani, mis annab vähima vigade arvu:

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

Toodud variantide proovimise järjekord kehtib seisuga 1. november 1998. Varem avaldatud artiklites on ta mõne üksiku variandi järjekorra poolest erinev.

4.4 Sõnajärg

On väidetud, et "Eesti keeles ei ole võimalik välja tuua statistilist põhisõnajärjemalli" ja et "Eesti sõnajärje ALUSEKS ei ole mitte süntaktilise struktuuri, vaid infostruktuuri printsiibid" (Tael 1988).

Samal ajal teame, et 40-50% tekstis ettetulevatest sõnadest on morfoloogiliselt mitmeti tõlgendatavad, mis tähendab seda, et sõnade süntaktilisi suhteid lauses on morfoloogiliste tunnuste abil raske kindlaks teha.

Seega on tegemist võimatuna näiva olukorraga: eestikeelse lause süntaktilist struktuuri ei saa justkui määrata ei sõnade järjekorra ega morfoloogiliste tunnuste alusel.

Paradoksi lahendus võiks olla järgmine. Lause süntaktilise struktuuri määravad ikkagi morfoloogilised tunnused; nende ühene tõlgendamine omakorda ei sõltu aga lause põhisõnajärjest, vaid sõnade lokaalsest kontekstist, mis on küllalt fikseeritud sõnajärjega. Antud väidet näib kinnitavat Markovi varjatud mudeli (MVM) kasutamise eksperiment eesti keele ühestamisel (Kaalep, Vaino 1998).

Osutub, et eesti keele puhul piisab mitmeti tõlgendatavate sõnade puhul 80% juhtudel ainult sõnavormi esinemise tõenäosusest ja sõnade järjestuse arvestamisest 2-3 sõnalises kontekstis, et õigesti otsustada, milline morfoloogilise analüüsi variant mitmest võimalikust valida.

5. PRAKTILISED TÖÖVAHENDID

Käesolevas töös käsitletakse kaht keeletehnoloogilise arvutitarkvara esindajat: morfoloogilist analüsaatorit ja ühestajat.

Morfoloogiline analüsaator on arvutiprogramm, mis mingi sõnavormi puhul võib määrata selle sõna algvormi, sõna struktuuri (formatiivid) ja morfoloogilise informatsiooni (nt. sõnaliigi, käände või pöörde, arvu jms). Erinevad morfoloogilised analüsaatorid erinevad üksteisest nii selle poolest, millist informatsiooni ja millise detailsusega nad väljastavad, kui ka selle poolest, milliseid meetodeid nad kasutavad.

Morfoloogiline *ühestamine* seisneb morfoloogiliselt analüüsitud lause igale sõnale tema võimalike morfoloogiliste märgendite hulgast õigete 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_ //*

5.1 ESTMORF, eesti keele morfoloogiline analüsaator

Eesti keele morfoloogia-analüsaator on praktilistest töövahenditest, mis uurimuse tulemusena loodud, kesksel kohal. Olles ise lingvistiline töövahend, on ta aluseks ka mitmetele kommertsrakendustele (nt. speller). Morf. analüsaator on töövahend, ilma milleta oleks raske ette kujutada ka automaatseid vahendeid keelekasutuse ja süntaksi uurimise tarvis.

ESTMORF on arvutiprogramm suvalise eestikeelse teksti analüüsimiseks. Teda saab kasutada nt. Interneti kaudu (http://www.filosoft.ee/html_morf_et/). ESTMORF on realiseeritud nii, et jooksvas tekstis olevaid sõnesid võrreldakse sõnastikus olevate lekseemide kombinatsioonidega. Võrdlemisel ei kasutata 2-tasemelisi reegleid (Koskenniemi 1983).

ESTMORFi peamised omadused on järgmised:

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

5.2 Ühestaja

Ühestaja on töövahend, mille vajalikkus saab selgeks niipea, kui morfoloogiline analüsaator on olemas. Nt. morfoloogiliselt analüüsitud, kuid ühestamata teksti alusel saab teha statistilisi uurimistöid ainult küllalt piiratud ulatuses; ka sagedussõnastiku koostamine on sel puhul peaaegu sama raske kui lihtsalt puhta teksti alusel.

Laialdaselt kasutatakse ühestamiseks statistilisi meetodeid. Eesti keele jaoks on praegu realiseeritud üks klassikalisi statistilisi ühestajaid - Markovi Varjatud bigramm-mudel (MVM). Statistiline ühestaja koosneb tegelikult kahest poolest: programmist ja keelespetsiifilistest andmetest, nn. keelemudelist. Programm tugineb üldtuntud algoritmile, on universaalne ja keelest sõltumatu. See, mis teeb statistilise ühestaja konkreetse keele jaoks kasutatavaks, on just keelemudel.

MVM puhul ei ole keelemudel midagi muud kui 3 ühestamisel kasutatavat tõenäosuste tabelit. Et neid kirjeldada, tuleb esmalt defineerida mõned mõisted. Esiteks peame silmas, et ühestamine kui omaette etapp teksti töötlemisel võib kasutada samu märgendeid sõnade morfoloogilise iseloomu kajastamiseks kui morfoloogiline analüsaator, aga ei pruugi. Tavaline ongi, et ühestaja jaoks defineeritakse omaette märgendite süsteem koos teisendusalgoritmiga, mis morfoloogilise analüsaatori märgendid teisendab ühestaja omadeks ja tagasi. Defineerime märgendite hulga $M = \{m_1\, m_2 \ldots m_n\}$, kus m_i on üks märgend. Mitteüheses tekstis võib ühel sõnal olla mitu märgendit; selle sõna märgendite komplekti nimetame mitmesusklassiks. Erinevate mitmesusklasside hulk $V = \{v_1\, v_2 \ldots v_q\}$ on loomulikult hulga M osahulkade hulk ($v_i \subseteq M$).

Tõenäosuste tabelid on järgmised:

1. Tõenäosuste vektor $E = \{e_1\, e_2 \ldots e_w\}$, kus e_i on tõenäosus, et m_i on lauses esimene märgend.
2. Maatriks $P = \{p_{kl}\}$, kus p_{kl} on tõenäosus, et märgendile m_k eelneb märgend m_l.
3. Maatriks $X = \{x_{kl}\}$, kus x_{kl} on tõenäosus, et mitmesusklassi v_l kuuluvatest märgenditest tuleb valida märgend m_k

Keelemudeli koostamiseks on teada mõned üldised printsiibid, kuid täpseid eeskirju mitte. On teada, et kuna MVM "näeb" ainult märgendeid ja nende tõenäosusi, siis märgendite süsteemi valik on peamine, mis eristab head MVM-ühestajat halvast. Samas ei ole olemas häid eeskirju, kuidas märgendite süsteemi teha; see on niivõrd keelespetsiifiline. Samuti on teada, et oluline osa statistilise ühestaja keelemudeli loomisel on treenimisel kasutatavate tekstide iseloom ja hulk: mida paremini vastavad tekstid tüüpilisele keelekasutusele ja mida rohkem neid on, seda parem.

Praegu kasutab statistiline ühestaja 88 ühestamismärgendit (ÜM), mis on valitud järgmiselt. Eristatakse omadussõnu, põhiarvsõnu, järgarvsõnu, nimisõnu, pärisnimesid, isikulisi asesõnu, muid asesõnu, lühendeid, verbe,

alistavaid ja rinnastavaid sidesõnu, hüüdsõnu, ees- ja tagasõnu, määrsõnu, punktuatsioonisümboleid ja tundmatuid sõnu.

Käändsõnade puhul eristatakse 5 käänet: nimetavat, omastavat , osastavat , lühikest sisseütlevat e. aditiivi ja "kõiki muid". Isikuliste asesõnade puhul eristatakse lisaks ka kolme isikut. Ei eristata ainsust ja mitmust.

Verbide puhul eristatakse kokku 13 ÜM-i: "ei", "ära", esimene pööre, teine pööre, kolmas pööre, kaudne kõneviis, "pole" ja "polnud", da-infinitiiv, 0-lõpuline vorm, tingiva kõneviisi vormid, käskiva kõneviisi vormid, ma-infinitiivi vormid, partitsiibid. Ei eristata ainsust ja mitmust ega aega.

Ühestaja tõenäosuste tabelid on saadud, treenides teda G. Orwelli "1984" eestikeelse tõlke peal (Orwell 1990), v.a. Lisa, mille suurus oli 75 000 sõna.

Ühestaja töö kvaliteeti iseloomustab tabel 8, mis on saadud ühestaja testimisel Vello Lattiku raamatu "Mihklipäeval. Mihklikuul" (Lattik 1983) 2005-sõnalisel väljavõttel:

	Alguses	Pärast ühestamist
Sõnu kokku	2005	2005
Tõlgendusi kokku	3450	2052
Mitteüheseid sõnu	852	47
Keskmiselt tõlgendusi sõna kohta	1,72	1,02
Mitteüheste sõnade protsent	42,49%	2,34%
Vale märgendiga sõnade protsent	0,1%	6,7%

Tabel 8. MVM bigramm-ühestaja töö kvaliteet.

Tabelis toodud arvud käivad morfoloogiliste märgendite (MM) kohta, mitte ühestamismärgendite (ÜM) kohta. See, et pärast ühestamist jääb osa sõnu mitmeseks, on seletatav sellega, et enne ühestamist võetakse mitu MM kokku üheks ÜMiks, kusjuures ühe sõna erinevad MMid teisenduvad tavaliselt siiski erinevateks ÜMideks. Kuid juhul, kui sõna erinevad MM teisenduvad üheks ÜMiks, MMide ühestamist ei toimugi. Nt. sõna *olema* on ainus verb, mille ainsuse ja mitmuse 3. pööre on homonüümsed – *on*. Kuna meie oma ÜMide valikul praegu ainsust ja mitmust ei erista, siis *on* jääb ühestamisest kõrvale.

Võrdlus hoopis teistel alustelt lähtuva ühestamismeetodiga – kitsenduste grammatikaga (Puolakainen 1997) näitab, et MVM esialgsed tulemused eesti keele peal ei jää praktiliselt alla ühestajale, mis kasutab inimese poolt formuleeritud konteksti arvestamise reegleid.

Võrdlus teiste keeltega, kus on kasutatud statistilisi meetodeid ühestamisel, näitab, et eesti keelele sobib antud meetod umbes sama hästi kui näiteks rootsi keelele, kus erinevus inimese ja arvuti poolt ühestamisel oli algul samuti 7% (Källgren 1996).

6. KOKKUVÕTE

Keeletehnoloogiline arendustöö, s.t. keeletehnoloogiliste toodete (nt. morfoloogilise analüsaatori ja spelleri) loomine on tihedalt seotud keeleressursside (nt. tekstikorpuste ja elektrooniliste sõnastike) kasutamisega. Juhul, kui keeleressursse pole, nõuavad keeletehnoloogia vajadused nende loomist. Keeletehnoloogilise arendustöö ja keeleressursside loomise käigus tuleb lahendada mitmeid teoreetilisi probleeme; nii varemtuntuid kui uusi.

Keeleressursid, keeletehnoloogilised tooted ja teoreetilised probleemid on omavahel seotud. Keeleressursside põhjal luuakse tooteid, mida omakorda saab kasutada uute ressursside loomiseks. Keeleressursid on toodete loomise aluseks; keeletehnoloogiavajadused omakorda mõjutavad ressursside kogumist ning loomist. Keeletehnoloogia võimaldab keelt paremini uurida; kuid toodete loomisel kerkivad üles ka uued teoreetilised probleemid.

Dissertatsioonis kirjeldatakse tulemusi, mis on saadud eesti keele tehnoloogilises arendustöös ja sellega seotud valdkondades: keeleressursside loomisel ja teoreetiliste küsimuste lahendamisel.

1. Keeleressursside loomine. On loodud miljoni-sõnaline nn. eesti kirjakeele baaskorpus ja 80 000-sõnaline põhjalikult märgendatud korpus G. Orwelli "1984" põhjal; 130 000 kirjet sisaldav sõnade andmebaas ja 38 000-sõnaline keerulise struktuuriga morfoloogilise analüsaatori leksikon.
2. Teoreetilised küsimused. On uuritud morfoloogiliste kategooriate süsteemi, lühikese sisseütleva käände ja vokaalmitmuse kasutatavust, produktiivseid tuletisi ja liitsõnu ning sõnajärge. Uurimistulemused on rakendatud praktilistes töövahendites.
3. Praktilised töövahendid. On loodud morfoloogiline analüsaator, mis on m.h. mitmete kommertsprogrammide aluseks (nt. speller, poolitaja, lemmatiseerija) ja muude keeletehnoloogiliste vahendite hädavajalikuks etapiks (nt. morfoloogiline ühestaja, süntaksi analüsaator). Morfoloogilist analüsaatorit on kasutatud ka keeleressursside loomisel, nt. morfoloogiliselt märgendatud korpuse tegemisel. On loodud ka statistikal põhinev morfoloogiline ühestaja.

ABSTRACT

The dissertation describes Estonian language technology development in 1991-1998, which has been tightly connected with creating language resources and theoretical problems in computational linguistics.

Part one, the introduction, outlines the ways language resources, linguistic tools and theoretical work are dependent on each other. It also gives a short characterization of the articles, included in the dissertation.

Part two, the background, puts the dissertation in the context of language technology and computational linguistics. It also gives a short overview of Estonian language resources before 1991.

Part three describes Estonian language resources, in creating of which the author has played an important part. These resources are:

1. The 1-million word corpus of Estonian literary language at the University of Tartu
2. The 80,000-word corpus of G. Orwell's "1984", containing rich mark-up
3. The lexical database of Estonian word-forms, containing 130,000 entries and 45,000 base forms
4. The lexicon of the Estonian morphological analyzer ESTMORF, containing 38,000 base forms and having a complex structure.

Part four describes theoretical issues in conjunction with language technology development:

1. The question of suitable morphological categories for computational treatment of Estonian
2. The usage of certain case forms (forms of the vocal plural and the short illative case) in real texts
3. The issue of productive derivation and compounding
4. Word order in conjunction with morphological disambiguation

Part 5 describes practical linguistic tools, creating of which has triggered new explorations in theoretical and practical computational treatment of Estonian: the morphological analyzer ESTMORF and a Hidden Markov Model disambiguator.

In the conclusion, the results achieved in Estonian language technology development, resource collection and theoretical explorations in conjunction of the former two are briefly outlined.

KIRJANDUS

Atkins, S., Clear, J., Ostler, N. (1992). *Corpus Design Criteria*. Literary and Linguistic Computing, kd 7, nr 1, lk 1-15.

Bel N., Calzolari N., Monachini M. (eds.) (1995). *Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets*. MULTEXT Deliverable D1.6.1B, Pisa.

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H-J., Petkevic, V., Tufis, D. (1998) *Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European languages*. COLING-ACL '98, Proceedings of the Conference, Kd. 1, lk. 315-319

Eesti Keele Grammatika (1995) *1*. Toim. M. Erelt; Eesti TA EKI, Tallinn.

Erjavec, T. (ed.) (1997) *Sample Corpus Collection and Preparation.* MULTEXT-East Final Report, D2.1 F http://nl.ijs.si/ME/CD/docs/mte-d21f/index.html

Erjavec, T., Ide, N., Petkevic, V., Véronis, J. (1996). *Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages*. Proceedings of the First European TELRI Seminar: Language Resources for Language Technology, lk. 87-98.

Erjavec, T., Lawson, A., Romary, L. (eds) (1998) *East Meets West — A Compendium of Multilingual Resources*. TELRI Association

Erjavec, T., Monachini, M. (eds.) (1997). *Specifications and Notation for Lexicon Encoding*. MULTEXT-East Final Report, D1.1. http://nl.ijs.si/ME/CD/docs/mte-d11f/index.html

Hein, I. (1994). *Practical realisation of the morhological analysis*. Viks, Ü. (toim.) Automatic Morphology of Estonian 1. Research Report. EKI, Tallinn, lk. 29-35

Hennoste, T. (1996). *Tartu University Corpus of Written Estonian: A Survey of the Structure of Texts and Principles of Selection*. H. Õim (toim.) Estonian in the Changing World. Tartu, lk. 7-32

Ide, N. (ed.) (1996): *Language-Specific Resources.* MULTEXT-East Intermediate Report, D1.2. http://www.lpl.univ-aix.fr/projects/multext-east/MTE2.html

Ide, N., J. Véronis. (1994). *MULTEXT (Multilingual Tools and Corpora)*. Proceedings of the 14th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan 1994, lk. 90-96.

Ide, N., Priest-Dorman, G., Véronis, J. (1996). *Corpus Encoding Standard*. http://www.cs.vassar.edu/CES/

Kaalep, H-J. (1996) *ESTMORF, a Morphological Analyzer for Estonian*. Kogumikus H. Õim (toim.) Estonian in the Changing World. Tartu, lk. 43-98

Kaalep, H-J. (1997) *An Estonian Morphological Analyser and the Impact of a Corpus on Its Development*. Computers and the Humanities 31: lk. 115-133

Kaalep, H-J. (1998) *Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator*. Keel ja Kirjandus 1/1998, lk 22-29

Kaalep, H-J., Prillop, R., Ehasalu, E. (1998). *The Role of Internet in Creating, Financing and Integrating Language Resources*. Proceedings of the First

International Conference on Language Resources and Evaluation, Granada, Kd. 2, lk 1149-1152
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
Kasik, R. (1984) *Eesti keele tuletusõpetus: õppevahend eesti filoloogia ja žurnalistikaosakonna üliõpilastele. 1. Substantiivituletus.* TRÜ, Tartu.
Kasik, R. (1992) *Eesti keele tuletusõpetus: õppevahend eesti filoloogia ja žurnalistikaosakonna üliõpilastele. 1. Adjektiivi- ja adverbituletus.* TRÜ, Tartu.
Kask, A. (1967) *Liitsõnad ja liitmisviisid eesti keeles*. Eesti keele grammatika 3.1., Tartu, 1967
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
Kull, R. (1967) *Liitnimisõnade kujunemine eesti kirjakeeles.* Dissertatsioon filoloogiakandidaadi kraadi saamiseks. ENSV TA KKI, Tallinn.
Källgren, G. (1996) *Linguistic Indeterminacy as a Source of Errors in Tagging*. COLING-96 proceedings, Copenhagen, 2. kd, lk 676-680.
Lattik, V. (1983) *Mihklipäeval. Mihklikuul*. Eesti Raamat, Tallinn, lk. 4-10.
Monachini M., N. Calzolari (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.
Muischnek, K. (1998) Korpused ja nende kasutamine. Magistritöö, Tartu
Orwell, G. (1990) *1984*, tlk. Elias Treeman. Loomingu Raamatukogu, Perioodika, Tallinn.
Peebo, J. (1997) *Eesti keele muutkonnad*. Tartu Ülikooli Kirjastus, Tartu.
Priest-Dorman, G., Erjavec, T., Ide, N., Petkevic, V. (1997): *Corpus Markup*. MULTEXT-East Final Report, D2.3 F http://nl.ijs.si/ME/CD/docs/mte-d23f/index.html
Puolakainen, T. (1998) *Eesti keele kitsenduste grammatika morfoloogiline ühestaja*. Keel ja Kirjandus 1, lk. 37-46
Sperberg-McQueen, C.M., Burnard, L. (eds.) (1994). *Guidelines for Electronic Text Encoding and Interchange*. Kd. I-II. ACH, ACL, ALLC. Chicago and Oxford.
Tael, K. (1988) *Sõnajärjemallid eesti keeles (võrrelduna soome keelega).* Preprint KKI-56, Tallinn.
Valgma, J., Remmel, N. (1970) *Eesti Keele Grammatika*. Valgus, Tallinn
Viks, Ü. (1992) *Väike vormisõnastik I. Sissejuhatus & grammatika; II. Sõnastik & lisad*. Tallinn.
Vuotilainen, A., Heikkilä, J., Anttila, A. (1992) *Constraint Grammar of English. A Performance-Oriented Introduction*. Univ. of Helsinki, Dept. of General Linguistics, No. 21
Yli-Vakkuri, V. (1993) *Tutkimushanke Itämeren piirin kielten kieliopillinen vertailu.— Studia comparativa linguarum orbis Maris Baltici*. Yli-

Vakkuri, V. (toim.) Studia comparativa linguarum orbis Maris Baltici I. Tutkimuksia syntaksin ja pragmasyntaksin alalta. Turku, lk. 9-12.

ELULOOKIRJELDUS

Heiki-Jaan Kaalep
Kodakondsus: Eesti
Sündinud: 19. mail 1962 Tallinnas
Abielus, 2 last
Aadress: Vaba 19, Tartu
Telefon: (27) 375 942
E-mail: hkaalep@psych.ut.ee

Haridus
1969-1980 Tallinna 44. Keskkool
1980-1985 Majandusküberneetika eriala TRÜ-s
1985-1988 TRÜ statsionaarne aspirantuur
1992 TÜ informaatikamagister

Erialane enesetäiendus
Õppevisiit Stockholmi Ülikooli arvutilingvistika osakonda tutvumaks korpuste alal tehtava tööga ja arvutilingvistikas kasutatava tarkvaraga Stockholmis, Rootsis 16 - 22. mai 1994
Õppevisiit Helsingi Ülikooli üldkeeleteaduse osakonda tutvumaks korpuste alal tehtava tööga ja arvutilingvistikas kasutatava tarkvaraga Helsingis, Soomes 31. oktoober - 5. november 1994
Osavõtt LSP-uurimise suvekoolist Gillelejes, Taanis, 14-22. juuni 1995

Erialane teenistuskäik
1985-1991 TRÜ tehisintellekti labori teadur
1991-1992 TÜ tehisintellekti labori juhataja
sept. 1993 - märts 1994 0,5 lektor TÜ majandusteaduskonna majandusinformaatika ja -modelleerimise instituudis
al. 1992 teadur TÜ eesti filoloogia osakonna üldkeele- teaduse õppetooli juures

Teadustegevus
Arvutilingvistika (morfoloogia, korpuslingvistika, keeletehnoloogia); 23 publikatsiooni.

CURRICULUM VITAE

Heiki-Jaan Kaalep
Citizenship: Estonian
Born on May 19th, 1962 in Tallinn
Married, 2 children
Address: Vaba 19, Tartu
Telephone: (27) 375 942
E-mail: hkaalep@psych.ut.ee

Education
1969-1980 School No. 44, Tallinn
1980-1985 The State University of Tartu, economic cybernetics
1985-1988 Postgraduate student of the State University of Tartu
1992 Master degree in Informatics (MSc)

Special courses
Study visit to the Dept. of Computational Linguistics of the University of Stockholm, Sweden on May 16 - 22, 1994
Study visit to the Dept. of General Linguistics of the University of Helsinki, Finland on October 31st - November 5th, 1994
Nordic Summer School on Research in LSP in Gilleleje, Denmark, June 14-22, 1995

Professional employment
1985-1991 researcher of the Laboratory of Artificial Intelligence of the State Univ. of Tartu
1991-1992 head of the Laboratory of Artificial Intelligence of the Univ. of Tartu
Sept. 1993 - March 1994 0.5 lecturer of the Institute of Economic Informatics and Modeling of the Univ. of Tartu
al. 1992 researcher of the Dept. of General Linguistics of the Univ. of Tartu

Scientific work
Computational linguistics (morphology, corpus linguistics, language technology); 23 publications.