Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1

Arvutuslingvistikalt inimesele

Tartu 2000

Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1

Arvutuslingvistikalt inimesele

Toimetaja Tiit Hennoste

Tartu 2000

Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1
Arvutuslingvistikalt inimesele
Toimetaja Tiit Hennoste
Kujundaja Roosmarii Kurvits

ISSN 1406–619X
ISBN 9985–4–0152–2
Tartu Ülikooli kirjastus
Tiigi 78, Tartu 50410
Tellimus nr. 399

Sisukord

Eessõna

Siinne kogumik avab Tartu Ülikooli üldkeeleteaduse õppetooli toimetiste sarja. Vormiliselt võib kogumiku teema tunduda ootamatu, aga sisuliselt on see loomulik. Arvutuslingvistika töörühm on töötanud TÜ üldkeeleteaduse õppetooli juures viimase asutamisest alates. Rühma juured lähevad tagasi vähemalt 1980. aastatesse, kui TRÜs töötas tehisintellekti labor, mille töö põhiteemaks oli keele mõistmise modelleerimine. Asjast huvitatuid võib juhatada venekeelse kogumike sarja "Trudy po iskusstvennomu intellektu" juurde, mis ilmus TRÜ Toimetistena (esimene 1978). 1980. aastate lõpus ühendati tehisintellekti labor vastloodud eesti keele laboriga, ja kui ülikooli reformimise käigus laborid varasemas mõttes kaotati, jäigi järele arvutuslingvistika uurimisrühm.

Vahepeal muutus põhimõtteliselt ka uurimisüksuse töö temaatika. Senise suhteliselt abstraktse ja teoreetilise temaatika asemele tuli eesti keele arvutitöötluse temaatika, ja esimene loomulik ülesanne oli eesti kirjakeele arvutikorpuse loomine. Täna ei kujuta ükski eesti keelt uuriv üliõpilanegi oma tööd ilma korpuseta ette, aga kümmekond aastat tagasi oli korpuse mõiste Eestis praktiliselt tundmatu.

Sellest said alguse uued tööd, ühelt poolt need, mis olid vajalikud korpuse kasutamiseks, ja teiselt poolt need, mis ainult korpuse olemasolu korral on võimalikud ja vajalikud muudeks rakendusteks.

Järgmine põhimõtteline pöördepunkt oli 1994/95. aastal, kui Eestil avanes võimalus osaleda Euroopa Liidu projektides COPERNICUS-programmi raames. Käibele tuli termin keeletehnoloogia, mille all mõistetakse arvutuslingvistika rakenduslikku väljundit, aga mille konkreetseks eesmärgiks on lahendada paljukeelse Euroopa suhtlemisprobleemid. Esimestest projektidest alates on eesti arvutuslingvistid osalenud enam kui 10 sellises projektis. Suur osa neist tulemustest kajastub siinseski kogumikus. On enam kui selge, et Eesti vajab oma eesti keele tehnoloogiat, et olla võrdväärne osaline tuleviku infoühiskonnas.

Uurimisrühma tööd on mitmel erineval viisil toetanud Eesti Teadusfond oma grantide kaudu, Teaduskompetentsi Nõukogu

sihtfinantseeringute kaudu, aga ka näiteks Avatud Eesti Fond arvutuslingvistika õpetusele antud toetuste kaudu. Tõhusat toetust on saadud ka HESPilt (*Higher Education Support Program*), eriti arvutuslingvistika õpetusprogrammi, aga ka raamatukogu koostamisel. 1997. aastast alates toimib Eesti keeletehnoloogia sihtprogramm, mis on esimene samm taotletava infoühiskonna suunas.

Rühma töö on viimastel aastatel laienenud mitmes suunas (morfoloogia, süntaks, semantika). Olulise uue alana tahaksin esile tõsta eesti suulise kõne uurimist, mis algas taas korpuse loomisest ja mis avab eesti keeleteaduses üldse uue uurimissuuna. On aga lisandunud ka praktilist eesmärki teenivaid projekte, nagu näiteks Zürichi ja Tartu ühisprojekt Webextrans.

Samas tahaksin mainida, et eelmistest aegadest on säilinud näiteks dialoogi arvutimudelitega tegelemine – see on teema, mis on peale 1980. aastaid taas muutunud üliaktuaalseks seoses võimalusega luua automaatse kõnetuvastuse ja -sünteesiga (so. ilma inimese vahenduseta) süsteeme suhtlemaks telefonitsi mitmesuguste andmebaasidega.

Tuleb rõhutada ka seda, et arvutuslingvistika, millele käesolev kogumik on pühendatud, ei ole algusest peale olnud ainult Tartu ülikooli tegevusala. Selle arendamisele Eestis on sama tõhusalt kaasa aidanud Eesti Keele Instituut (varasem Keele ja Kirjanduse Instituut) ja Küberneetika Instituut, ehkki neis on arvutuslingvistika lähted olnud teised kui Tartu ülikoolis.

Siinne kogumik ei ole organiseeritud kindla teema ümber, vaid tema eesmärgiks on anda läbilõige hetkel Eesti arvutuslingvistikas (keeletehnoloogias) toimuvast. Samas on see sihilikult koostatud nii, et Tartu Ülikoolis arvutuslingvistika eriala üliõpilased – ja vajadusel muidugi ka teiste alade üliõpilased – saaksid kirjutisi kasutada õpingutes lektüürina.

Kogumikus on esindatud neli eesti arvutuslingvistika keskust: Tartu ülikooli üldkeeleteaduse õppetool ja arvutiteaduse instituut, Eesti Keele Instituut ning Küberneetika Instituut.

Kogumiku algusosa moodustavad artiklid, milles tegeldakse eesti kirjakeele arvutimorfoloogia mitmesuguste probleemidega. Siia kuuluvad Ülle Viksi, Heli Uibo, Tiina Puolakaineni ja Heiki-Jaan Kaalepi – Tarmo Vaino artiklid. Kogumiku teise kontsentri moodustavad mitmesugused leksikaalse info arvutianalüüsi probleemid. Sellesse rühma kuuluvad Margit Langemetsa, Kadri Videri jt, Heili

Orava ja Neeme Kahuski jt artiklid. Sellele järgnevad artiklid, mis tegelevad eesti kirjakeele korpuse ja selle süntaktilise märgendamisega (Tiit Hennoste – Kadri Muischnek ning Kadri Muischnek jt). Neljanda rühma moodustavad artiklid, milles tegeldakse suulise kõne ja suhtlemise probleemidega. Siia kuuluvad Tiit Hennoste jt, Mare Koidu – Haldur Õimu ja Meelis Mihkla jt artiklid.

Haldur Õim
Tartu Ülikooli üldkeeleteaduse professor

Eesti keele avatud morfoloogiamudel*

Ülle Viks
Eesti Keele Instituut

1. Mis on mis

1.1. Avatud morfoloogiamudel

EE määratleb **mudelit** kui objekti, mis on kindlas vastavuses mingi teise objektiga (originaaliga), asendab seda tunnetamisel ja võimaldab selle kohta saada vahendatud andmeid. Mudelit kasutatakse siis, kui originaali otsene uurimine on võimatu või raskendatud. Inimkeel kuulub kahtlemata selliste objektide hulka, mida saab uurida eelkõige mudelite abil. Järgnevas käsitluses on juttu ühest eesti keele olulisemast allsüsteemist – **morfoloogiast** – ning selle modelleerimisest ja kirjeldamisest.

Avatud morfoloogiamudel on avatud kahes mõttes. Esiteks tähendab mudeli avatus seda, et tema rakendusulatus ei ole piiratud mingi kindla hulga sõnadega, vaid et süsteem saab hakkama ka talle tundmatute sõnadega – nii nagu inimene oskab käänata-pöörata ka neid sõnu, mida ta kunagi varem kasutanud pole. Inimese morfoloogiapädevusse kuulub teatud hulk sõnavorme, mida ta tunneb ja oskab lauses kasutada, kuid lisaks neile kindlasti ka teatud hulk **reegleid**, mille abil ta on võimeline täiesti uusi sõnavorme moodustama (või neist aru saama).

Teiseks tähendab mudeli avatus seda, et tema üksikute allsüsteemide elemente on võimalik muuta-täiendada, ilma süsteemi ennast ümber tegemata – nii nagu inimene õpib juurde uusi sõnavorme ja reegleid (või korrigeerib olemasolevaid), sobitades neid olemasolevasse süsteemi. Mehhanismid, mille abil inimene oma morfoloogilist pädevust realiseerib, on püsivamad kui konkreetsed reeglid või sõnavormid.

1.2. Morfoloogiasüsteem

Igas süsteemis on oluline välja tuua: 1) üksused ja nendevahelised seosed ehk süsteemi struktuur ja 2) protsessid, mis on aluseks

* Projekti toetab praegu Eesti Teadusfond (grant nr 3862).

süsteemi funktsioneerimisele. Täpsustan nende mõistete sisu avatud morfoloogiamudeli raames (mis ei tähenda, et mõni teine mudel ei võiks samu asju teises valguses näha).

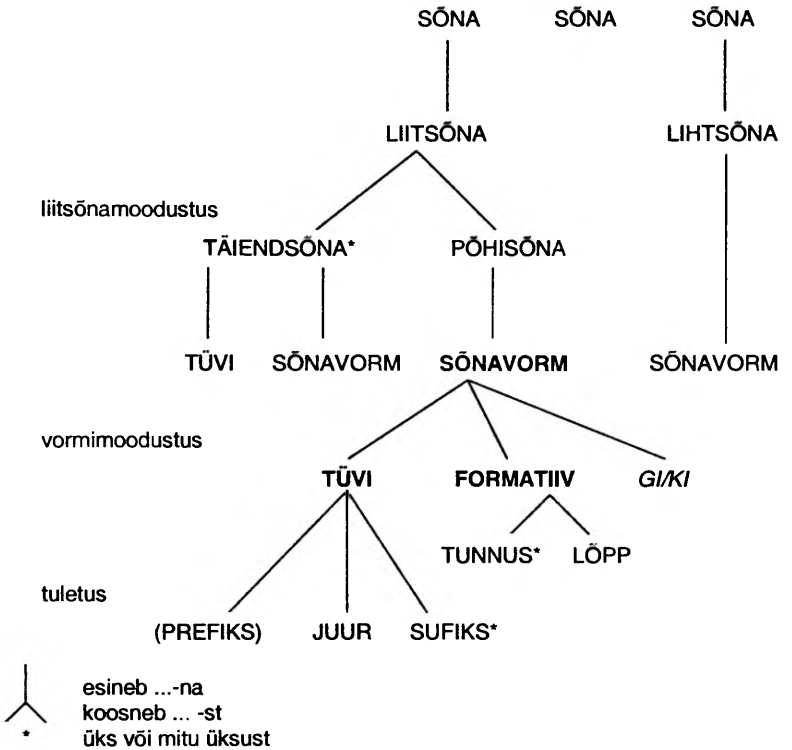
1.2.1. Üksused

Morfoloogiasüsteemi keskne üksus on **sõnavorm** (ehk lihtsõnavorm), mis koosneb morfeemidest. Leksikaalne morfeem (ehk tüvi) kannab sõnavormi leksikaalset tähendust ja grammatilised morfeemid (ehk tunnused ja lõpud) kannavad sõnavormi grammatilisi tähendusi, nt *hammas[te[ga* → 'hammas' & 'pl' & 'kom' Eesti morfoloogia on suures osas aglutinatiivne, kuid palju on ka fusioonijuhtumeid, kus piirid morfeemide vahel on hägusad. See pole aga siinse käsitluse teema. Avatud morfoloogiamudel piirdub minimaalse morfoloogilise liigendusega, mis jagab sõnavormi kaheks: **tüveks** ja **morfoloogiliseks formatiiviks**. Morfoloogiline formatiiv kannab seega kogu morfoloogiliste tähenduste komplekti sõltumata sellest, kas teda saab edasi liigendada või mitte, või kas igal grammatilisel tähendusel on oma fonoloogiline realisatsioon või mitte, nt *hammas[tega* → 'hammas' & 'pl kom' *hamba[ga* → 'hammas' & 'sg kom'

Tüvi võib edasi alluda tuletusliigendusele, mis jagab tüve juureks ja tuletusmorfeemi(de)ks (nt *mõist|mine*). Lihtsõnavormist kõrgemal tasandil on liitsõnaliigendus, mis jagab liitsõna täiendsõna(de)ks ja põhisõnaks. Täiendsõnu võib olla mitu (kuni 4, nt *all+maa+raud+tee+jaam*) ning täiendsõnana võib esineda kas seotud tüvi (*nais+koor, liihi+laine, tõmb+lukk*) või muutevorm (*jalg+ratas, lehe+külg, õige[ks+mõist|mine, kä[tel+kõnd* jne). Minimaalsel morfoloogilisel liigendustasandil käituvad nii tuletised kui liitsõnad komplekssete tüvedena, millele võivad liituda formatiivid (*õigeksmõistmise[ga, kätelkõnni[st*). Peaaegu iga sõnavormi lõpus võib mingis kontekstis esineda veel liide *gi/ki* (*minu[st_ki, ärka[b_ki, õige[ks+mõist|mise[ga_gi*).

Eesti keele sõnavormide üldstruktuuri iseloomustab järgmine skeem, kus iga liigendustasand esindab omaette grammatilist allsüsteemi: liitsõnamoodustus, vormimoodustus ehk morfoloogia, tuletus ehk derivatsioon. Avatud morfoloogiamudeli keskne osa on vormimoodustus, kuid sellega liituvad ka tuletus ja liitsõnamoodustus.

Sõnavormi üldstruktuur:



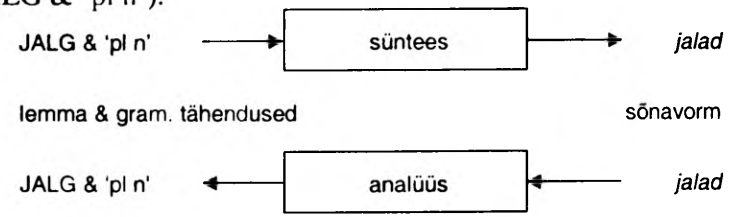
Näidetes kasutatavad tähised:

&	ja
[	morfoloogilise formatiivi piir
+	liitsõnapiir
|	tuletusliite piir
\	fusiivse formatiivi piir
_	*gi/ki* eelne piir

1.2.2. Protsessid

Morfoloogiasüsteemi kesksed protsessid on sõnavormide süntees ja analüüs. Morfoloogilise **sünteesi** sisend on algvorm ehk lemma ja sõnavormi grammatilised tähendused (nt JALG & 'pl n') ning väljund on vastav sõnavorm (*jalad*). Morfoloogilise **analüüsi** sisend ja väljund on vastupidised: sisendiks on sõnavorm (*jalad*) ning väljundiks on lemma koos vastavate grammatiliste tähendustega

(JALG & 'pl n').



Sünteesi käigus valitakse sisendi elementidele vastavad üksused: lemmale JALG vastav tüvevariant *jala* ning grammatiliste tähenduste komplektile 'pl n' vastav morfoloogiline formatiiv *d*. Seejärel ühendatakse need üksused sõnavormiks *jalad*. Analüüsi käigus liigendatakse sõnavorm potentsiaalseteks üksusteks *jala [d* ja leitakse tüvevariandile *jala* vastav lemma JALG ning formatiivivariandile *d* vastav grammatiliste tähenduste komplekt 'pl n'

Esitatud skeem on nii üldine, et kehtib kõigi morfoloogiasüsteemide puhul. Erinev on ainult see, mis toimub sünteesi ja analüüsi kastide sees. See võib olla lihtne otsing sõnastikus, aga võib olla ka väga keeruline paljude sõnastike ja reeglite võrk.

Analüüsi- ja sünteesiprotsesse võib vaadelda pööratavatena ainult teatava lihtsustuse korral. Protsessid, mis toimuvad sisendi ja väljundi vahel, on kummagi suuna puhul paljuski erinevad. Kuid ometi on tegemist ühe ja sama morfoloogilise allsüsteemiga. Analüüs ja süntees kasutavad samu andmeid ja neid protsesse on võimalik selliselt arvutis realiseerida, et nad kasutavad mitmeid ühiseid programmimooduleid. Seetõttu saab neist ka koos rääkida.

1.2.3. Algvorm

Algvormi valik sõltub morfoloogiamudeli iseloomust ja kirjelduse eesmärkidest ning algvorme võib ka eesti morfoloogiakirjeldustes leida mitmesuguseid, nt

- üks reaalne vorm (sg nominatiiv, sg partitiiv; supiin, imperatiivi preesens, ...);
- mitu reaalset vormi (sg nominatiiv ja genitiiv, *da*-infinitiiv ja indikatiivi preesens, ...);
- üks kunstlikult konstrueeritud vorm, mis sisaldab reeglite tööks vajalikku infot, nt *liik(e/me)* sõna *liige : liikme* jaoks.

Mõeldav on kasutada eri allsüsteemide kirjeldustes erinevaid algvorme: tüvemuutusreeglite jaoks ühtesid, muuttüüpide käsitlemisel teisi. Nii ei saa aga toimida juhul, kui eesmärgiks on ühtne funktsioneeriv mudel. Avatud morfoloogiasüsteemi algvormiks on valitud sõnastikes kasutatav märksõnavorm ehk **lemma**: noomenil singulari nominatiiv ja verbil supiin (*ma*-infinitiiv), sest see annab parima võimaluse siduda omavahel grammatikat ja sõnastikku.

1.3. Sõnastik, grammatika ja erandid

Grammatika ja sõnastik on põhilised komponendid, mis mingil kujul on olemas kõigis morfoloogiasüsteemide kirjeldustes – nii traditsioonilistes kui automaatsetes.

1.3.1. Sõnastik

Sõnastik kujutab endast **üksuste** loendit, mis sisaldab allsüsteemi seisukohalt vajalikke andmeid iga üksuse kohta. Morfoloogia-sõnastikku kuuluvad:

- leksikaalsed üksused ehk sõnad (lemmad) ja nende tüved (või tüvevariandid), mille juurde kuuluvad andmed iseloomustavad sõna üldisi morfoloogilisi omadusi, aga ka konkreetse tüvevariandi käitumist sõnavormides (paradigmas, tekstis);
- grammatilised üksused ehk morfoloogilised formatiivid (või formatiivivariandid), mille puhul on oluline anda neile vastavad grammatiliste tähenduste kombinatsioonid;
- sõnavormid koos morfoloogilise infoga.

1.3.2. Grammatika

Grammatika sisaldab **reegleid**, mis kirjeldavad üksuste seoseid ja funktsioneerimist süsteemis. Morfoloogiasüsteemis on kõige olulisemad:

- struktuurireeglid (kombinatoorika),
- protsessireeglid (teisendused),
- tuvastusreeglid (tingimused).

Struktuurireeglid kirjeldavad üksuste kombinatoorikat kahel eri tasandil:

- sõnavormi tasandil näidatakse süntagmaatilisi seoseid üksikutes sõnavormides,

- paradigma tasandil esitatakse paradigmaatilisi seoseid sõnade kõigi vormide vahel (paradigmas).

Reegli üldkuju on 'A & B' ja see näitab, milline üksus millisega koos esineb ja millises järjestuses. Alljärgnevast kuuluvad struktuurireeglite hulka morfotaktika ja allotaktika reeglid.

Protsessireeglid kirjeldavad üksustega toimuvaid tegevusi ja muutusi süsteemi funktsioneerimisel, st morfoloogilise sünteesi ja analüüsi käigus. Reegli üldkuju on 'A → B' ja see näitab, mis millega asendatakse. Alljärgnevast kuuluvad protsessireeglite hulka tüvemuutuste reeglid, lisanäiteks võiks olla transkriptsioonivõi teisendused.

Tuvastusreeglid kirjeldavad tingimusseoseid üksuste või nende omaduste ja tegevuste vahel. Reegli üldkuju on 'kui A, siis B' ja see näitab, mis tingimusel miski esineb või toimub. Tuvastusreeglid kuuluvad sageli muude reeglite koosseisu, nt protsessireegli rakendumistingimuste määramiseks, aga neil on ka iseseisev staatus: tuvastusel põhinevad kitsenduste grammatikad, mis esitavad eri tasandite piiranguid. Alljärgnevast kuuluvad tuvastusreeglite hulka üksustevaheliste piiride tuvastus (silbitus, liitsõnapiirid) ja sõna klassikuuluvuse tuvastus (tüüp, sõnaliik).

1.3.3. Erandid

Lisaks reeglitele kuuluvad iga loomuliku keele grammatika juurde ka erandid, mis oma esitusviisi poolest on sõnastikunähtused (sõnavormid). Ükski loomulik keel ei ole täiesti reeglipärane, sest pika arengu käigus on paljud erinevad tegurid jätnud temasse oma jälgi. Uutest nähtustest rääkides ja teiste keelte mõjul tuleb keelde palju uusi sõnu, kuid teatud nihked tekivad aja jooksul ka grammatilises süsteemis.

Erandit saab defineerida **reegli kaudu**: erand on objekt, mille puhul reegli rakendumistingimused on küll täidetud, kuid mis ometi ei allu reeglile. Nt sõna *t'ank* võib nimetada häälikumuutusreegli k→g/n_V suhtes erandiks, sest näidatud kontekstis (*n* ja vokaali vahel) peaks *k* asenduma *g*-ga (nagu toimub sõnades *l'ank : langi* või *t'ankima : tangib*), aga sõnas *t'ank* jääb *k* püsima ja nõrgeneb ainult välde (*t'ank : tanki*).

Erand on suhteline mõiste: erandite hulk grammatikakirjelduses tuleneb otseselt sellest, kui palju ja kui keerulisi reegleid

fikseeritakse. Kui taotleda väga täpseid reegleid ja ammendavat kirjeldust, võib eranditest täiesti loobuda – reegel võib ju kehtida ka üheainsa sõna puhul (sel juhul tuleb tema rakendumistingimusena esitada terve sõna). Kui aga taotleda lihtsaid reegleid ja ülevaatlikku kirjeldust, on ka erandeid rohkem (lähemalt vt Viks 1997).

Erandlikud on keeles kõige tavalisemad ja sagedamini kasutatavad vormid nendest sõnadest, mis kuuluvad üldkasutatava põhisõnavara hulka, nt *on (olema), läks (minema), mehed (mees), häid (hea)*. See hulk on aga oma suhtelisusest hoolimata piiratud: erandeid on võimalik meelde jätta. See aga, mida harva kasutatakse, peab olema reeglipärane. Muidu pole võimalik seda suhtluses kasutada: kõik sõnavormid ei mahu mällu. Eriti oluline on reeglipärasus uute sõnade puhul: inimesed oskavad ka tundmatust sõnast vorme moodustada ja suudavad mõistatada võõra vormi jaoks algvormi (et seda näiteks sõnastikust otsima minna).

1.3.4. Sõnastiku ja grammatika integratsioon

Morfoloogiasõnastik ja grammatika on väga tihedalt seotud: nad kirjeldavad üht ja sama nähtust – morfoloogilist süsteemi –, kuigi teevad seda erineval viisil. Sõnastik sobib paremini individuaalse info jaoks, grammatika sobib paremini üldistuste jaoks, kuigi esitatav ainestik ise on sama. Sõnastik sisaldab iga üksuse kohta andmeid, mis on vajalikud reeglite (grammatika) tööks, grammatika sisaldab reegleid, mis võimaldavad üksusi omavahel siduda. Kui kaks üksust on omavahel seostatavad reegli vahendusel, siis pole tarvidust hoida neid mõlemaid sõnastikus. Piisab, kui sõnastikus on neist üks ja teine saadakse sellest reegli abil. Seega annab grammatika ja sõnastiku sidumine võimaluse vähendada sõnastiku koormust ning muudab keelekirjelduse kokkuvõttes ülevaatlikumaks ja üldisemaks. Mida suurem on reeglite osakaal, seda vähem andmeid on vaja esitada sõnastikus.

Grammatika ja sõnastik üldjuhul ei asenda teineteist, kuid suurem osa morfoloogilisest süsteemist on võimalik ära kirjeldada kas ühe või teise vahenditega. Kumb on ülekaalus, sõltub konkreetse keelekirjelduse suunitlusest. Arvutimudel ei vaja andmete dubleerimist, seevastu inimestele mõeldud käsitlused – nii grammatikad kui sõnastikud – ei saa läbi ilma liiasuseta.

Avatud morfoloogiamudel üritab integreerida grammatika ja sõnastiku erinevad võimalused maksimaalse ökonoomsusega,

vältides kirjelduse liiasust. Pearõhk on grammatikal: kõik see, mis on morfoloogias üldine ja reeglipärane, esitatakse formaalsete reeglite kujul, so. grammatikana, ja ainult need nähtused, mis reeglite alla ei mahu, esitatakse erandite loendina sõnastikus. Seda võib pidada reeglipõhiseks morfoloogiaks üsna äärmuslikul kujul.

2. Grammatika allsüsteemid eesti morfoloogias

Kui morfoloogiasüsteem oleks üles ehitatud ainult sõnastikule – täiesti ilma reegliteta – siis vajaks ta sõnastikku, mis sisaldab kõiki sõnavorme koos täieliku morfoloogilise infoga. **Sõnastik 1** on kasutatav ilma reegliteta:

Lemma	SL	Gr.täh	Sõnavorm
JALG	S	sg n	j'alg
JALG	S	sg g	jala
JALG	S	sg p	j'alga
JALG	S	sg ill	jala[sse
...			
JALG	S	pl n	jala[d
JALG	*S*	*pl g*	*j'alga[de*
JALG	*S*	*pl g*	*j'alg\e*
JALG	S	pl p	j'alga[sid
JALG	S	pl p	j'alg\u
JALG	S	pl ill	j'alga[desse
...			
SÜDA	S	sg n	süda
SÜDA	S	sg g	südame
SÜDA	S	sg p	südan[t
SÜDA	S	sg ill	südame[sse
...			

Kogu morfoloogiline analüüs ja süntees taanduvad sellisel juhul sõnastikuotsingule: *JALG 'S' & 'pl g' –j`algade, j`alge* (2 kirjet).

Selline sõnastik oleks hiigelsuur, sest eesti keeles on palju muutevorme. Noomenil on 14–15 käänet ainsuses ja mitmuses, lisaks hulgaliselt paralleelvorme mitmuses. Verbil on ainuüksi lihtvorme ligi 60, lisaks hulk analüütilisi aja- ja eitusevorme. Keskeltläbi on igal eesti sõnal 33 muutevormi (koos liitega *gi/ki* 66). Samas ei saa sõnavormide sõnastikku kunagi ammendavaks pidada, sest

uusi sõnu saab alati juurde teha: tuletada, liita, laenata, välja mõelda jne. Ja igaühel neist on taas oma muutevormid. Seega eesti keele jaoks puhas sõnastikupõhine morfoloogiasüsteem ei sobi.

Edasi vaatame, kuidas erinevate reeglite lisamine aitab vähendada sõnastiku mahtu. Reeglid, mis kirjeldavad morfoloogiasüsteemi struktuuri seaduspärasusi ja süsteemis toimuvaid protsesse, jagunevad oma liigi ja rakendussfääri alusel mitmeks allsüsteemiks, millest igaühel on ka omad erandid ja oma info, mida süsteem oma tööks vajab. Põhilised allsüsteemid (reeglimoodulid), mis on olulised eesti morfoloogia jaoks, on järgmised: morfotaktika, allotaktika, tüvemuutused, tüübituvastus (vt ka Viks 1994), liitsõnapiiri tuvastus.

2.1. Struktuurireeglid: morfotaktika

Kõige universaalsem morfoloogia allsüsteem on morfotaktika. Morfotaktika reeglid määravad ühelt poolt ära kogu keele vormimoodustussüsteemi struktuuri: grammatiliste kategooriate valiku ja realiseerumisvõimalused ning sõnaklassidele vastavad morfoloogilised paradigmad (nt verbi- ja noomeniparadigma). Teiselt poolt määravad morfotaktika reeglid ära sõnavormi sisemise struktuuri: millised üksused millises järjestuses ja mis tingimustel võivad ühes sõnavormis koos esineda, nt noomenil `tüvi & arv & kääne`. Et avatud morfoloogiamudel grammatilisi morfeeme ei lahuta, siis sama struktuur esineb siin kujul `tüvi & arv_kääne`. Morfotaktika reeglid laienevad ka tuletistele ja liitsõnadele ning määravad ära sõnamoodustuses osalevate morfeemide kombinatoorika võimalused.

Vormimoodustusse puutuvat morfotaktilist infot esindab sõnastikus **sõnaliigi** tähis, millele vastavate paradigmade kirjeldused ise on alati väljaspool sõnastikku – grammatikas. Paradigma kirjeldus näitab, millistest liikmetest paradigma koosneb ning millised grammatiliste tähenduste kombinatsioonid on muutevormide (ehk paradigma liikmete) taga. Iga grammatiliste tähenduste komplektiga seatakse vastavusse tema fonoloogiline realisatsioon – morfoloogiline formatiiv.

Morfotaktika reeglite erandid on näiteks sõnad, mille puhul teatud grammatilised kategooriad jäävad realiseerimata, nt *mõlema* (sg nom puudub), pronoomen *iga* (mitmus puudub), *pidama* 'kohustatud olema' (impersonaal ja käskiv kõneviis puuduvad). Või sõnad, millel on mõni täiendav paralleevorm, nt sõnadel *jalg, silm, rind* on

mitmuse omastavas lisavormid, mida teistel sõnadel ei ole: *jalge, silme, rinde.*

Ainuüksi morfotaktika reeglite abil oleks täiesti võimalik kirjeldada morfoloogiasüsteemi funktsioneerimist (sõnavormide sünteesi ja analüüsi) aglutinatiivsete keelte puhul, kus on küll rohkesti muutevorme, kuid kus üksused ei varieeru. Sel juhul peaks sõnastikus olema lemma koos sõnaliigiga (mis määrab paradigma valiku). Ja reeglid seavad igale grammatilisele tähendusele vastavusse sobiva morfeemi (nt 'pl g' – *de*). Lemma ja morfeemide kokkupanekul saadaksegi sõnavormid.

2.2. Struktuurireeglid: allotaktika

Et eesti keeles võivad varieeruda kõik morfoloogilised üksused, nii tüved kui ka morfoloogilised formatiivid, siis ei piisa nende kombinatoorika kirjeldamiseks morfotaktika reeglitest, mis töötavad üksuste tasemel. Üksused võivad küll anda korrektse morfotaktilise struktuuri, kuid üksuste variandid võivad olla omavahel sobimatud. Nt struktuur `tüvi & 'pl g'` on igati õige. Sõnal LIIGE on kaks tüvevarianti: *liige* ja *l'iikme*, mitmuse omastava formatiivil on variandid: *te* ja *de*. Seega oleks kokku neli erinevat võimalust variante kombineerida. Korrektne mitmuse omastava vorm selle sõna jaoks on aga ainult üks – *l'iikme[te* (tugev aste & *te*), vrd *k`iike[de, hõige[te, kolge[de.*

Õige valiku tegemiseks on vaja kontrollida variantide omavahelist sobivust. Reegleid, mis kirjeldavad üksusevariantide (allomorfide) valiku ja koosesinemise tingimusi sõnavormis, olen nimetanud morfotaktika eeskujul **allotaktika** reegliteks (morfotaktika – morfeemide kombinatoorika, allotaktika – allomorfide kombinatoorika).

Kõige ökonoomsem viis allotaktika reeglite esitamiseks on **morfoloogiline klassifikatsioon**, kus muuttüüp esindab üht võimalikku tüvevariantide ja formatiivivariantide allotaktilist asetust paradigmas: millised variandid millistes muutevormide esinevad. Puhtal kujul allotaktika reegleid esindab VVS-i (Viks 1992) klassifikatsioon, mis liigitab kogu sõnade hulga kolme morfoloogia seisukohalt olulise tunnuse järgi:

1) formatiivivariantide komplekt paradigmas,
2) tüve astmevaheldusmallid,
3) tüve lõpuvaheldusmallid.

Vaheldusmalle on kaks liiki vastavalt tüvemuutuste eri liikidele (astmemuutus ja lõpumuutus, vt lähemalt 2.3.), ja nad iseloomustavad tüvevariantide paigutust paradigmas. Vaheldusmallid kuuluvad klassifikatsiooni liigitusaluste hulka, sest nad on seotud otseselt konkreetsete paradigmaliikmetega ja määravad selle, millises muutevormis millist tüvevarianti kasutada, kuid (NB!) mitte seda, kuidas ühest tüvevariandist teist moodustada. Näiteks järgmised erinevate astmevaheldusmallidega sõnad alluvad kõik samale astmevaheldusliku tüvemuutusele, vrd (tähised allpool):

sg n	sg g	sg p	pl g
v'aat (t)	vaadi (n)	v'aati (t)	v'aati[de (t)
põder (n)	põdra (n)	p`õtra (t)	p`õtra[de (t)
vaade (n)	v`aate (t)	vaade[t (n)	vaade[te (n)
mõõde (n)	m`õõtme (t)	mõõde[t (n)	m`õõtme[te (t)

Samuti võib ka üks lõpumuutus olla seotud erinevate lõpuvaheldusmallidega, vrd

sg n	sg g	sg p	pl g
r'audne (a)	r'audse (b)	r'audse[t (b)	r'audse[te (b)
hobune (a)	hobuse (b)	hobus[t (c)	hobus[te (c)
hammas (a)	h'amba (b)	hammas[t (a)	hammas[te (a)
sipelgas (a)	sipelga (b)	sipelga[t (b)	sipelga[te (b)

Kolme eri tunnuse alusel saadud koondklassifikatsiooni klasse nimetatakse muuttüüpideks. Ühte muuttüüpi kuuluvad sõnad käituvad ühtmoodi kõigi aluseks võetud tunnuste järgi: neil on ühesugused formatiivivariandid kõigis muutevormides ning neil paiknevad tüvevariandid (nii astme- kui ka lõpumuutuse variandid) paradigmas ühtmoodi.

Allotaktilist infot esindab sõnastikus tüübinumber, mis juhatab grammatikas esitatud tüübikirjelduse juurde. **Tüübikirjeldus** kujutab endast paradigma põhivormide esitust sellisel kujul, mis fikseerib kõik tüübile omased variandikombinatsioonid: iga põhivormi jaoks tema tüvevariant (tüvekoodi kujul) koos konkreetse formatiivivariandiga ([järel). Tüübikirjeldus näitab ka võimalikku paralleelvormide kasutust (& paralleelvormide vahel) ning vormi puudumist tüübis (X vastavas positsioonis). Näide:

tp	sg n	sg g	sg p	sg adt	pl g	pl p	Näide
06	an[	at[	an[t	X	an[te	at[id	*vaade*
20	a0[	b0[	b0[	b0g[	b0[de	b0[sid	*tüvi*
22	at[	bn[	bt[	bt[	bt[de	bt[sid & btv\	*v`aat*
25	at[	bn[	bt[	bt[	bnv\ & bt[de	btv\ & bt[sid	*haril`ik*

Tüvevarianti esindab **tüvekood**, mille elemendid osutavad tüvevariandi liigile. Tüvekoodi elemendid on:

- lõpumuutuse järgi: a – lemmatüvi, b – muutetüvi, c – sekundaarne muutetüvi;
- astmemuutuse järgi: t – tugevaastmeline tüvi, n – nõrgaastmeline tüvi, 0 – astmemuutuseta tüvi;
- grammatiliste tüvemuutuste järgi: v – vokaalmitmuse tüvi, g – gemineerunud aditiivitüvi.

Nt tugevaastmelisele lemmatüvele vastab tüvekood at; nõrgaastmelisele vokaalmitmuse tüvele, mis baseerub muutetüvel, vastab tüvekood bnv; astmemuutuseta geminaattüvele, mis baseerub lemmatüvel, vastab tüvekood a0g jne.

Tüübikirjeldus annab tavaliselt juhised sõna põhivormide moodustamiseks. Muude paradigmasse kuuluvate vormide (nn analoogiavormide) moodustamine on seotud põhivormidega analoogiareeglite kaudu ega mängi rolli tüüpide eristamisel.

Kombinatoorikareeglite kasutamine muudab süsteemi sõnastiku oluliselt väiksemaks. Kõigi sõnavormide asemel on tarvis esitada ainult kõik tüvevariandid koos oma koodidega. Lisaks tuleb sõnastikku panna aga morfotaktika ja allotaktika reeglite erandid (täiendavad paralleelvormid, erandliku tüvega vormid jms).

Allotaktika reeglite (ehk klassifikatsiooni) erand ei ole päris ühemõtteline nähtus, sest tüübistik ise ei ole ühemõtteline. Et liigitusaluseid on mitu ja nad on erilaadsed, siis ka saadav klassifikatsioon on tegelikult mitme allklassifikatsiooni ühendus (Viks 1977). Tüübierand on sõna, millel mõni vorm tervikuna või teatud vormide tüved, või teatud vormide formatiivid erinevad sellest, mida allotaktika reegel (tüübikirjeldus) selle tüübi jaoks ette näeb. Suurem osa erandliku sõna muutevormidest moodustatakse täiesti regulaarselt. Nt sõna *olema* on *tulema*-tüübi (tüüp 36) erand, sest tal on ainult kahes muutevormis teistsugune kuju *`on*–'Ind Pr Sg 3' ja 'Ind Pr Pl 3' (vrd *`on – tule[b, tule[vad*). Kõik muud vormid on tal

sarnased *tulema*-tüübi vastavate vormidega (*ole[n – tule[n, `oll[a – t`ull[a, ol[nud – tul[nud* jne). Sõna *ajama* on erandlik tüübi *elama* (27) suhtes, sest tal on impersonaalis irregulaarne tüvevariant *'ae* (*'ae[takse* jne, vrd *ela[takse* jne).

Sõnastik 2, millega koos töötavad morfotaktika ja allotaktika reeglid, võiks olla selline:

Lemma	Tüüp_SL	Tüvekood	Tüvevariant	Vormierandid
VAAT	22_S	at	v'aat	
VAAT	22_S	bt	v'aati	
VAAT	22_S	bn	vaadi	
VAAT	22_S	btv	v'aate	
VAAT	22_S	bnv	vaade	
JALG	22_S	at	j'alg	
JALG	22_S	bt	j'alga	
JALG	22_S	bn	jala	
JALG	22_S	btv	j'algu	
JALG	22_S	bnv	jalu	
				pl g: & j'alge
				pl ab: & jaluta
SÜDA	04_S	a0	süda	
SÜDA	04_S	b0	südame	
				sg p: südan[t
HINNE	06_S	an	hinne	
HINNE	06_S	at	h`inde	
LUUSTIK	25_S	at	l`uust`ik	
LUUSTIK	25_S	bt	l`uust`ikku	
LUUSTIK	25_S	bn	l`uustiku	
LUUSTIK	25_S	btv	l`uust`ikke	
LUUSTIK	25_S	bnv	l`uustike	
NUUSTIK	02_S	a0	nuustik	
NUUSTIK	02_S	b0	nuustiku	

Morfotaktika ja allotaktika reegleid rakendades on juba võimalik eesti keele automaatne morfoloogia tööle panna. Nii on VVS-i baasil realiseeritud EKI varasemad analüüsi- (Hein 1994) ja sünteesiprogrammid (Kuusik 1994). VVS-i allotaktika reeglistikku (lisatud on liitsõnareeglid ja osa tuletusreegleid) kasutab ka Filosofti morfoloogiline analüsaator (Kaalep 1996), mis töötab Microsofti

Wordi eesti keele spelleris. Kuid morfotaktika ja allotaktika reeglitega piirduv süsteem on veel suletud süsteem, mis suudab sünteesida ja analüüsida ainult neid sõnu, mis on sõnastikus olemas ning on varustatud reeglite tööks vajalike andmetega (sõnaliik, tüübinumber ja võimalikud tüvevariandid koos tüvekoodidega). Iga uus sõna tuleb sõnastikku lisada – ja see nõuab lisatööd inimeselt, kes peab määrama muuttüübi ja sõnaliigi, ning moodustama kõik vajalikud tüvevariandid ja määrama nende tüvekoodid.

2.3. Protsessireeglid: tüvemuutused

Esimene samm avatud süsteemi poole on tüvemuutusreeglite rakendamine (Kuusik 1995, 1996). Tüvemuutuste reeglid moodustavad mitu erinevat allgrammatikat – vastavalt tüvemuutuse liigile. Põhilisi tüvemuutusi on kahte liiki:

- sisemuutus ehk astmemuutus,
- lõpumuutus.

Sise- (ehk astme)muutuse puhul seob reegel tüve tugevaastmelist (t) ja nõrgaastmelist (n) varianti. Astmemuutus avaldub sageli ainult välte- (ehk aktsendi)muutuses (*m'etsa : metsa*), kuid sellega võivad kaasneda ka mitmed tüve sisehäälikute muutused, nt klusiili teisenemine (*saade : s'aate*), assimilatsioon (*k'anda : kanna*), konsonandi kadu (*'ulgu[ma : ulu[b, jõge : j'õe*) jne.

Lõpumuutuse puhul seob reegel sõna algvormi tüve e lemmatüve (a) ja muutetüve (b), mis esineb teistes vormides. Lõpumuutusi on väga erinevaid, nt tüvevokaali lisandumine (*isand : isanda, k'and : k'anda*), vokaali teisenemine (*suvi : suve, jõgi : jõge*), konsonandi kadu (*sipelgas : sipelga*), häälikujärjendi asendumised (*kolmas : kolmanda, puder : pudru, vestle[ma : vestel[da*) jne.

Mõlemat liiki tüvemuutused on jälgitavad eraldi tüvepaarides (nagu ülaltoodud näidetes), kuid võivad esile tulla ka korraga ühes ja samas tüvevariantide paaris, nt *pääse : p'ääsme* (lõpumuutus e → me, astmemuutus 2 → 3), *kinnas : k'inda* (lõpumuutus s → 0, astmemuutused 2 → 3 ja nn → nd), *v'aatle[ma : vaadel[da* (lõpumuutus le → el, astmemuutused 3 → 2 ja t → d).

Põhilised tüvemuutused ise ei ole otseselt seotud kindlate grammatiliste tähendustega. Üks ja sama tüvevariant esineb tavaliselt mitmes erinevas muutevormis, millel on erinevad grammatilised tähendused, nt *hobuse : hobuse[ga : hobuse[d, hobus[t :*

hobus[tega jne. Üks ja sama tüvemuutusreegel võib toimida noomeni- ja verbimorfoloogias, aga ka derivatsioonis, vrd *link : lingi – lonki[ma : longi[b – vanker : vangerda[ma.* Variantide seost konkreetsete muutevormidega iseloomustab tüvevaheldus, mida kirjeldavad vaheldusmallid (vt 2.2.).

Tüvemuutuste allsüsteemi on siinses mudelis lülitatud ka kaks grammatiliste tähendustega seotud fusiivse tüvemuutuse liiki, mille korral grammatiline tähendus ei realiseeru selgelt eristatava aglutinatiivse formatiivina, vaid tüvemuutusena. **Vokaalmitmuse** tüve (v) moodustamisel teiseneb ainsusetüve lõpuvokaal (nn tüvevokaal), nt *kavaleri : kavalere, l'ille : l'illi, haril'ikku : haril'ikke, vana : vanu.* **Aditiivi geminaattüve** (g) moodustamisel pikeneb (või tugevneb ja pikeneb) tüve lühike sisekonsonant (kusjuures kaasneb ka sekundaarne aktsendimuutus 1 → 3), nt *maja : m`ajja, lume : l'umme, tuba : t'uppa, jõge : j`õkke.*

Omaette rühmadena on vormistatud ka mõningad **sekundaarsed** tüvemuutused, nagu:

- morfonoloogilise distributsiooni nähtused (tähis r), nt tüvelõpu *i* asendumine *e*-ga formatiivialgulise *i* ees (*osuti & id → osute[id*), pika tüvelõpuvokaali lühenemine *i*-ga algava formatiivi ees (*id'ee & id → id`e[id*);
- teatud ortograafiateisendused, mis rakenduvad mõne muu reegli järel ja korrigeerivad tulemust, nt konsonantühendi reegel (*kuppel : kuppli → kupli*).

Iga tüvemuutuste liik on kirjeldatav omaette reeglite komplektiga, mis moodustavad formaalse allgrammatika, kusjuures kummagi tüvemuutuste suuna jaoks (nt tugev → nõrk ja nõrk → tugev) on eraldi allgrammatika. Grammatika tähis avab ühtlasi ka tema olemuse, nt

G_tn:	tugev aste (t) → nõrk aste (n)	*v'aati : vaadi*
G_nt:	nõrk aste (n) → tugev aste (t)	*vaadi : v'aati*
G_ab:	lemmatüvi (a) → muutetüvi (b)	*v'aat : v'aati*

Tüvemuutuste allgrammatika koosneb osaliselt järjestatud ümberkirjutusreeglitest (*rewriting rules*). Reegli üldkuju on x → y / z_q, st järjend x asendatakse järjendiga y, juhul kui talle eelneb järjend z ja järgneb järjend q. Suurtäht reeglis tähistab häälikuklassi, mis on reeglite jaoks ära defineeritud, väiketäht on tema ise. Nt reegel t → d / N_V tähendab, et asendus *t → d* toimub kontekstis, kus *t* ees on

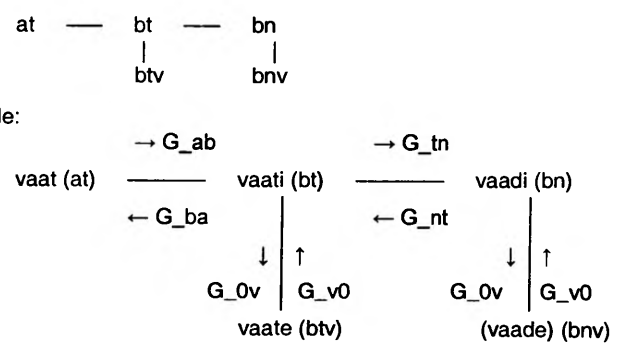
klassi N kuuluv täht (N=lnr) ja *t* järel on vokaal (klass V). Sisend *kaarti* sobib reegli tingimustega ja väljundiks on tüvevariant *kaardi*.

Näiteks üks fragment grammatikast G_tn (häälikuklassid: V=aeiouõäöü; Q=kptfš; L=lmnr; J=jv):

Q1Q1 → Q1 / !VV_V	lõppu → lõpu, bluffi → blufi
ss → s / !VVL_V	valssi → valsi
ss → s / VV_V	poissi → poisi
p → b / VV_V	vaapa → vaaba
p → b / VL_V	kulpi → kulbi
p → b / V_JV	lupja → lubja

Kui sisendi kuju vastab reegli tingimustele, aga näidatud muutust ei toimu, siis see tüvevariantide paar on antud reegli suhtes erand. Näiteks kui grammatikas G_ab on reegel 0 → i /VVt_#–(# tähistab tüve lõppu), siis reeglipäraselt käituvad *i*-tüvelised sõnad *v`aat : v`aati, p`eet : p`eeti, radikul'iit : radikul'iiti, sal'uut : sal'uuti* jne, selle reegli suhtes erandid on aga muude tüvevokaalidega sõnad *l'aat : l'aata, l'iit : l'iitu, `eit : `eite* jne. Või kui grammatikas G_nt on reegel d → t /ee_i, siis reeglipärased on *peedi : p`eeti, ankeedi : ank`eeti, epiteedi : epit'eeti* jne, erandid on *pleedi : pl'eedi, logopeedi : logop`eedi, mopeedi : mop`eedi jne.* Kui selle reegli ette lisada tõkestav reegel d → d / opee_i, siis jääks erandlikuks ainult *pleedi : pl'eedi.*

Tüvevariantide moodustamise käiku suunab avatud morfoloogiamudelis **tüvejuht**, mis iga muuttüübi jaoks näitab ära, milline tüvevariant millisega on seotud. Seos on kahesuunaline, nii et sünteesi käigus on tüvejuhti järgides võimalik jõuda algvormi ehk lemma tüvest iga vajaliku vormi tüvevariandini, ja vastupidi, analüüsi käigus on võimalik jõuda iga konkreetse vormi tüvevariandist lemmatüveni. Nt tüübi 22 jaoks kehtestab tüvejuht järgmise skeemi:



See, milline formaalne allgrammatika läheb käiku mingi tüvevariandi saamiseks, tuletatakse automaatselt tüvekoodide põhjal. Kui on vaja moodustada at-tüvest bt-tüvi (*v'aat* → *v aati*), siis valitakse tüvekoodide erinevate elementide järgi grammatika G_ab (lemmatüvest muutetüvi). Kui on vaja moodustada btv-tüvest bt-tüvi (*v'aate* → *v aati*), siis rakendub grammatika G_v0 (mitmusetüvest ainsusetüvi).

Tüvemuutusreeglite rakendamiseks ei ole vaja täiendavat sõnastikuinfot, sest reeglite rakendumistingimused on reeglis endas fonoloogilise kontekstina antud.

Sõnastikku peavad jääma aga erandid – need tüvevariandid, mida allgrammatikates antud reeglite abil ei saa moodustada või mis moodustuksid valesti. Igal tüvemuutuste allgrammatikal on omad erandid, mis vaadatakse läbi enne reeglite poole pöördumist, ja ühel sõnal võib osa tüvevariante olla erandlikud, osa reeglipärased. Kõige rohkem erandeid annavad tüvevokaali lisamise reeglid allgrammatikas G_ab (*laata, liitu, eite* jne), kuid küllalt palju aitab siin tuletussufiksite arvestamine reeglite rakendumistingimustes. Sufiksid eelistavad tavaliselt mitte-*i*-list tüvevokaali, nt *-kond : -konda, -us : -use, -ik : -iku, -v : -va, -m : -ma* jne, lihttüved eelistavad *i*-d. Suhteliselt palju erandeid annavad ka nõrgaastmelisest tüvevariandist tugeva astme moodustamise reeglid allgrammatikas G_nt. Eriti tülikad on mõned võõrliited, mis tugevas astmes on erinevad, kuid nõrgas astmes on samakujulised, nt *-iidi : 'iiti* ja *-iidi : 'iidi* (vrd *kloriidi : klor'iiti* ja *kloriidi : klor'iidi*). Sellisel juhul on reeglisse valitud see sufiks, mis on eesti keeles produktiivsem (kuigi ka selle üle otsustamine pole alati lihtne).

Sõnastik 3, millega koos töötavad lisaks eelmistele ka tüvemuutuste reeglid, võiks olla selline (vrd ka sõnastik 2):

Lemma	Tüüp_sõnaliik	Tüveerandid
VAAT	22_S	
JALG	22_S	ab_ba j'alg j'alga nt_tn jala j'alga
SÜDA	04_S	ab_ba süda südame
HINNE	06_S	
LUUSTIK	25_S	
NUUSTIK	02_S	

Tüvemuutusreeglite lülitamine süsteemi grammatikaossa annab võimaluse hoida sõnastikus iga reeglipärase sõna jaoks ühtainsat tüvekuju (lemmat). Sõnastikuinfoks jääb iga lemma juurde seega ainult sõnaliik ja tüübinumber. Kuigi sõnastik on veel olemas, väheneb sel teel süsteemi sõnastiku maht küllalt palju, sest tüvemuutustega on seotud u ¾ eesti keele sõnavarast, ja ühel sõnal on vähemalt kaks, aga sageli kuni viis tüvevarianti. Tüvemuutuste süsteemi realiseerimine arvutis näitas, et formaalsete reeglitega on võimalik kirjeldada tüvemuutusi u 90% ulatuses VVSi sõnavarast. Ainult 10% sõnadest on mingi tüvemuutusreegli suhtes erandlikud.

Kuigi kokkuhoid sõnastiku mahus on tuntav, pole süsteem ainult tüvemuutusreeglite abil veel sõnastikust vabaks saanud – sealt on tarvis leida iga sõna jaoks sõnaliik ja muuttüübi number, millest sõltub nii tüvemuutuste kui ka morfotaktika ja allotaktika allsüsteemide töö.

2.4. Tuvastusreeglid: tüübi- ja sõnaliigituvastus

Sõnastikust aitavad päris vabaks saada tuvastusreeglid, mille abil saab kindlaks teha sõna tüübikuuluvuse ja sõnaliigi. Sõna muutmisvõimalused eesti keeles sõltuvad suurel määral sellest, milline on sõna enda fonoloogiline struktuur: mitu silpi on algvormis ja muutetüves, kus on pea- ja kaasrõhk, millises vältes on tüvi, millised häälikud on tüve lõpus, millised on sisehäälikud jne. Kõiki morfoloogia jaoks olulisi tunnuseid saaks reeglite koostamisel arvestada sel juhul, kui sisend oleks morfonoloogilises transkriptsioonis, mis tähistab vähemalt aktsenti (nn kolmandat väldet) ja morfoloogilist rõhku. See eeldab aga sõnastiku olemasolu, või siis automaatset teisendust ortograafiast morfonoloogilisse transkriptsiooni (selle tegemist on EKIs pisut katsetatud).

Avatud morfoloogiamudel orienteerub tavalisele ortograafiale, et toime tulla suvaliste eestikeelsete sõnadega, mida süsteemisõnastikus pole. Seetõttu kasutatakse reeglites kahte tunnust, mis on igal juhul sõna ortograafilisest kujust automaatselt kättesaadavad: silpide arvu ja sõna lõpuhäälikuid (või -tähti). Silpide arvu sisendsõnas määrab silbitusprogramm, mis kasutab oma tuvastusreegleid ja erandeid. Lõpuhäälikute arv ei ole piiratud ja tegelikult võivad nende hulka olla kaasatud ka sõna sisehäälikud traditsioonilises mõttes. Konkreetsete tähtede kõrval kasutatakse reeglites häälikuklassi

tähiseid, kusjuures häälikuklassid on ühised tüvemuutusreeglites kasutatavate klassidega. Tuvastustunnuste kooslust reeglis olen nimetanud struktuurimalliks (vt ka Kuusik, Lind, Viks 1995).

Tuvastusreeglite sisendiks on sõna algvorm e lemma, mis esineb tavalistes leksikonides märksõnana ja mis on ka muidu sõnade nimetamisvormiks. Reeglite väljundiks on muuttüüp ja sõnaliik (VVSi klassifikatsiooni järgi). Reeglistik jaguneb kaheks põhiliseks allgrammatikaks: eraldi on verbireeglid ja noomenireeglid (viimastega koos on reeglid muutumatute sõnade jaoks). Neile eelneb väike reegliplokk, mis tuvastab, kas sisend on verb või mitte ja suunab sisendsõna edasi vastavalt sellele. Verbi aitab tuvastada lemmavormi (supiini) formatiiv *-ma*, mis edasise analüüsi jaoks kõrvaldatakse, nii et mõlemad grammatikad töötavad lemmatüvedega.

Reegli kuju on x y → z_q, mida tuleks lugeda nii: kui silpide arv tüves on x ja lõpuhäälikud vastavad järjendile y, siis kuulub sõna muuttüüpi z ja sõnaliiki q. Reegel 2 VV → 26_S ütleb, et kahesilbiline sõna, mille lemmatüve lõpus on 2 vokaali, kuulub tüüpi 26 ja on substantiiv, nt *fopaa*. Kui sõna võib muutuda kahe eri tüübi järgi, siis on väljundeid kaks ja nende vahel on kas tilde ~ (võrdväärsed paralleeltüübid) või küsimärk ? (esimene paralleltüüpidest on eelistatum), nt 3 ikkus → 11_S~09_S (nt *suutlikkus*) või 3 VVline → 12_A?10_A (nt *tõeline*).

Reeglid on järjestatud, nii et konkreetsemad struktuurimallid on eespool ja üldisemad tagapool. Tuvastusreeglite näiteks üks väljavõte noomenigrammatikast koos näidetega (V = vokaal, C = konsonant, Q = kpt):

2 Clik → 25_A	*petlik, kunstlik, piinlik*
2 CCnik → 25_S	*kunstnik*
2 VVCnik → 25_S	*üürnik*
2 VVstik → 25_S	*luustik*
2 CVQ → 02_S	*seelik, tehnik, rästik, taldrik, järsak, tulek, sõiduk, rätsep, vikat* jne

Tuvastusreeglite erandid on sõnad, mille fonoloogiline struktuur vastab reeglites kirjeldatud struktuurimallile, aga muuttüüp (või sõnaliik) on siiski teine. Näiteks '2 Clik' on produktiivne adjektiivimall, kuid sama struktuuriga on ka üksikuid substantiive (tüüp on sama), nt *ämblik, aadlik, puuslik*; struktuurimall '2 VVstik' määrab sõna enamasti tüüpi 25_S, aga sama struktuuriga sõnu leidub ka tüübis 02, nt *nuustik, päästik, kaustik.*

Kõige raskem on ära tunda muutumatuid sõnu, millel pole spetsiifilist tuletussufiksit, nt *all, nõnda, vähe, sest* jne. Ortograafilise sõnakuju puhul tekitab suuri raskusi pearõhu leidmine, millest sageli sõltub sõna tüübikuuluvus kõige rohkem, vrd *valang* (rõhk 1. silbil – tüüp 02) ja *volang* (rõhk 2. silbil – tüüp 22). Palju erandeid annavad 2-silbilised *e*-lõpulised sõnad, mis jagunevad mitme tüübi vahel, nt 01: *tüüne : tüüne*, 02: *homne : homse*, *rase : raseda,* 04: *tase: taseme*, 05: *ranne: randme*, 06: *vanne : vande*, 10: *vaene vaese*, 16: *kõne : kõne*. Suurem osa neist allub siiski reeglitele, millest kõige produktiivsem on 2 Cne → 02_A (nt *homne*). Kui lisada veel *ne-* ja *ke*-liited pika vokaali järel (reeglid 2 VVne → 10_A (nt *vaene*) ja 2 VVke → 12_S?10_S (nt *lõoke*) ja määrata ülejäänud Ce-sõnad tüüpi 06 (reegel 2 Ce → 06_S), siis on suurem osa 2-silbilisi *e*-sõnu saanud õige tüübimääratluse. Ülejäänud tuleb arvata erandite hulka.

Pärast tüübi- ja sõnaliigi tuvastusreeglite lisamist jäävad sõnastikku ainult erandid (vrd sõnastikud 2 ja 3), (H = lühike konsonant):

SÜDA	04_S	(reegel: 2 VHa → 17_S, nagu muda, reha, lisa, ...)
NUUSTIK	02_S	(reegel: 2 VVstik → 25_S, nagu luustik, ...)

Tüübituvastuse abil muutubki morfoloogiasüsteem avatuks. Kogu vajalik info süsteemi tööks tehakse kindlaks sõna enda fonoloogilise kuju põhjal ja ainult need sõnad, mille algvormi fonoloogiline struktuur ei võimalda antud reeglite abil tüüpi õigesti määrata (erandid), peavad jääma sõnastikku. Kokkuvõttes väheneb sõnastiku osa morfoloogiasüsteemis oluliselt ja grammatika osa suureneb.

Sõnu, mille ortograafiline algvorm ei sisalda piisavalt infot tüübi määramiseks, on VVS-i sõnade hulgas u 14%. Kui määrata ainult muuttüüp (ilma sõnaliigita), siis kahaneks erandite hulk 7–8%ni (vt Viks 1995b). Praegused reeglid on häälestatud VVSi klassifikatsioonile ja sõnaliigimääratlusele, aga kuna reeglid ise on vormistatud tekstifailina, siis sama programmi kasutades saab tuvastada tüüpe ka muude morfoloogiliste süsteemide järgi, kui reeglid ja erandid vastavalt ümber teha.

2.5. Morfoloogiline süntees ja analüüs

Avatud morfoloogiasüsteem funktsioneerib üldjoontes järgmiselt (Viks 1995a). Sünteesi puhul läbib sisendsõna (lemma) kõigepealt tuvastusmooduli, mille reeglid annavad väljundisse muuttüübi ja sõnaliigi. Järgneb tüvemuutuste moodul, kus genereeritakse tüvejuhti järgides kõik vastavas tüübis ettenähtud tüvevariandid (tüvemuutusreeglite abil). Sünteesi lõpetab vormimoodustuse (ehk kombinatoorika) moodul, kus allotaktika reeglite (tüübikirjelduse) järgi sobitatakse kokku nõutava muutevormi jaoks vajalik tüvevariant ja formatiivivariant (Kuusik, Viks 1998).

JALG & 'S' & 'pl g'

TUVASTUS	sõnaliik, tüüp
TÜVEMUUTUSED	tüvevariandid
KOMBINATOORIKA	sõnavormid

j'alga[de, j'alg\e

Analüüsi puhul kasutatakse neidsamu mooduleid, kuid teise algoritmi järgi. Sõnavormi liigendamisel (kombinatoorika moodul) järgitakse allotaktika reegleid, mille tulemusena saadakse iga eraldatud formatiivivariandi jaoks teatud valik võimalikke tüvevariantide koode. Igast liigendamisel saadud tüvevariandist moodustatakse tüvemuutuste moodulis kõik võimalikud lemmavormid. Lõpuks kontrollitakse tuvastusmoodulis, kas saadud lemmad sobivad vastava tüübi sõnaks.

jalgade

KOMBINATOORIKA	tüvi & formatiiv
TÜVEMUUTUSED	lemmad
TUVASTUS	tüüp, sõnaliik

JALG & '22_S' & 'pl g'

2.6. Sõnamoodustus

Sõnamoodustuse allsüsteemid – tuletus ja liitsõnamoodustus – kasutavad osaliselt samu reegleid mis vormimoodustus. Ühised on näiteks tüvemuutusreeglid ning tüübi ja sõnaliigi tuvastamise reeglid. Üksused, millega neis allsüsteemides opereeritakse, kuuluvad aga teisele liigendustasandile (vt 1.2.1.) ja seetõttu on ka kombinatoorikareeglid erinevad.

2.6.1. Tuletus

Tuletussüsteemi kombinatoorikareeglid sarnanevad vormimoodustuse allotaktika reeglitele, näidates tuletiste võimalikke struktuure: milline tüve- või juurevariant esineb koos millise tuletussufiksiga (või selle variandiga). Reegleid täiendavad rakendustingimused, mis esitavad eelkõige tuletusaluse sõnaliigipiiranguid, aga ka fonoloogilisi tingimusi. Näiteid:

- substantiivisufiks *mine* liitub verbi lemmatüvele lisatingimusteta (h'akka[ma V_29at: *h'akka* → *h akka|mine*, l'eid[ma V_34at: *l'eid* → *l'eid|mine*);
- substantiivisufiks *mus* liitub verbi lemmatüvele, kui selle lõpus on *ne*, *i* või *u* (paljune[ma V_27a0: *paljune* → *paljune|mus*, l'eppi[ma V_28at: *l'eppi* → *l'eppi|mus*, h'arju[ma V_27a0: *h'arju* → *h arju|mus*);
- adverbisufiks *lt* liitub adjektiivitüve samale variandile, mida vormimoodustuses kasutatakse sg genitiivi vormis (ablas A_07bt: *apla* → *apla|lt*, punane A_12b0: *punase* → *punase|lt*, k'urb A_22bn: *kurva* → *kurva|lt*);
- substantiivisufiks *us* liitub fusiivselt adjektiivi tugevaastmelisele või astmevahelduseta muutetüvele (ablas A_07bt: *apla* → *'apl|us*, *n'app* A_22bt: *n'appi* → *n app|us*, tore A_02b0: *toreda* → *tored|us*, kasul'ik A_25bt: *kasul'ikku* → *kasul'ikk|us*).

Tuletusmoodul töötab praegu ainult analüüsiprogrammi koosseisus. Tuletiste sünteesimiseks oleks vaja lisaks ka semantilisi piiranguid, nt tegijanime sufiks *ja* seostub aktiivset tegevust väljendavate verbidega (vrd *esineja*, aga mitte *pimeneja*). Ka analüüsi on kaasatud ainult kõige produktiivsemad ja formaalsemad tuletusreeglid, mis semantilisi piiranguid ei vaja või mille puhul nende puudumine vähem häirib. Semantiliste tingimuste formaliseerimine ei ole võimalik ilma eelneva uurimistööta.

2.6.2. Liitsõnapiiri tuvastus

Kõik vormimoodustuse reeglid töötlevad ainult lihtsõnu. See on eesti keele puhul ka täiesti põhjendatud, sest liitsõnade käänamisel muutub üksnes viimane komponent. Sellest leidub vaid üksikuid kõrvalekaldumisi: arvsõnade esikomponent muutub kolmes esimeses

käändes (nt *k`aks+kümmend – kahe+k`ümne – k`ahte+kümme[t – kahe+k`ümne[sse*) ja paarissõnadel käänduvad mõlemad pooled kõigis käänetes (nt *`emb-k`umb – emma-kumma – emba-k`umba, emma[l-kumma[l; `üks+ainus – ühe+`ainsa – `üht+ ainsa[t – ühe[st+ ainsa[st*). Üldiselt käändub liitsõna nii, nagu tema lõpukomponendile vastav lihtsõna, sõltumata sellest, mitu komponenti liitsõnas kokku on või millises vormis on täiendsõna(d), vrd *suusa+h`üppe+mägi* (nagu *mägi*) – *suusa+h`üppe+m`äe[le – suusa+h`üppe+mäge[sid – suusa+h`üppe+m`äkke; m`äkke+t`õus* (nagu *t`õus*) – *m`äkke+tõusu[ga – m`äkke+t`õus\e* jne.

Seega liitsõna lülitamiseks automaatsesse morfoloogiasüsteemi on vaja ainult teha kindlaks tema viimane komponent ning kõik muud reeglimoodulid tegelevad sellega nagu lihtsõnaga (mis võib olla ka tuletis). Liitsõnapiiri tuvastus reeglite abil ei ole aga lihtne ülesanne. See vajab mitut reeglikomplekti ja mitmeid abiloendeid.

Osaliselt aitavad sõnaosade piire leida **fonoloogilised** reeglid:

- teatud häälikuühendid saavad esineda ainult liitsõna piiril:
  - v+m *terav+meelne, operatiiv+mälu* (erand: *servmine*);
  - s+h *sinakas+hall, kordamis+harjutus* (erandid: *ishias, ekshibitsionism*);
- täpitähed esinevad eesti sõnades tavaliselt 1. silbis (va. mõned võõrsufiksid), see osutab liitsõnapiirile kusagil täpitähe ees: *ladva+õun, iga+päevane, kordus+trükk, maailmameistri+võistlused.*

Arvesse tulevad ka **statistilised** reeglid loendite kujul:

- teatud järjendid tõenäoliselt sisaldavad liitsõnapiiri: *aja+kir, aa+i, bi+el, äes+ol;*
- sagedasemad liitsõnakomponendid on võimalik ette anda: *maa+, +maa, pea+, +pea, eba+*

Sagedamate liitsõnakomponentide arvestamine toob enamasti kaasa liiga palju piire (*maa+lima* pro *maalima, no+maa[d* pro *nomaad, pea+mine* pro *peamine*), ja osa pakutud piire tuleb kõrvaldada. Selleks saab kehtestada keelureeglid. Kasulikud on näiteks:

- **fonoloogilised** piirangud: teatud ühendid ei esine eesti keeles sõna(osa) alguses või lõpus, nt *_km* või *km_* (nt *maa+*, aga *maak+mineraal*);

- **grammatilised** piirangud: teatud sõnaliigid ei esine liitsõna järelkomponendina (nt side- või kaassõna) või esinevad piiratult (nt verb või pronoomen).

Liitsõnaga seotud reeglistik ei ole EKI mudelis pole veel lõpule viidud (varasem versioon vt Hein 1995). Liitsõnapiiri otsingul tuleb ilmselt rohkem kasutada sõnastike abi, sest liitsõnamoodustuse reeglid ei saa läbi ilma leksikaalse semantikata – see allsüsteem keeles ei ole aga veel piisavalt formaliseeritud.

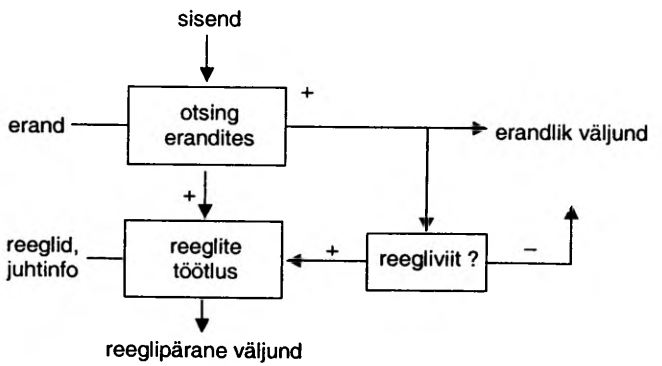
Liitsõnamoodul töötab praegu samuti ainult analüüsiprogrammis, mis arvestab liitsõnavõimalusega igas morfoloogilise liigenduse tsüklis. Sünteesi puhul peab liitsõnapiiri tuvastus toimuma esimeses järjekorras, sest tüübituvastuse reeglid vajavad viimase komponendi silpide arvu.

3. Avatud morfoloogiamudel arvutis

3.1. Tööpõhimõte

Reeglipõhise morfoloogia üldine tööpõhimõte, mis kehtib kõigi moodulite puhul, on järgmine. Sisendüksust otsitakse kõigepealt mooduli erandite sõnastikust. Kui otsitav üksus on seal olemas, siis saadakse sõnastikust ka vajalik väljund ja reeglid jäetakse vahele. Kui otsitavat üksust sõnastikust ei leita, siis suunatakse sisend reegliplokki, mille väljundiks on soovitav reeglipärane üksus. Reegleid töötlevad spetsiaalsed **reeglite interpretaatorid**.

Homonüümsete üksuste puhul võib üks homonüümidest olla reeglipärane, teine (teised) mitte. Sel juhul on sõnastikus erandi juures antud vastav viit, ning pärast erandliku vormi väljastamist suunatakse sama sisend edasi ka reegliplokki ja väljundeid saab mitu.



3.2. Nõuded tarkvarale

EKI tahab olla eelkõige teadusasutus, mitte tarkvarafirma. Seetõttu ei ole morfoloogiasüsteemi arendamisel võetud sihiks teha laiatarberakendust, vaid on arvestatud peamiselt uurimisteemadest tulenevaid vajadusi. Et tarkvara on EKI jaoks sekundaarne väljund, siis ei pööra me väga suurt tähelepanu selle tehnilistele parameetritele, vaid eelkõige sisulisele kvaliteedile: **pole tähtis et väljund tuleks kiiresti, vaid et ta oleks õige.**

Peamised nõuded, mida oleme silmas pidanud, on sellised:

- **keeleline adekvaatsus** ja usaldusväärsus: kõige olulisem on, et väljund oleks lingvistiliselt korrektne ja vastaks kehtivatele õigekeelsuse normingutele;
- **avatus**: süsteem peab ühelt poolt korrektselt töötlema kõiki eesti keele grammatika seisukohalt õigeid sõnu ja sõnavorme, sh ka tundmatuid (uusi sõnu, mitmesuguseid termineid jne), teiselt poolt peab süsteem võimaldama muuta ka keelereegleid (kui keel ise muutub);
- **paindlikkus**: süsteem (ja selle osad) peavad olema kasutatavad mitmetes erinevates rakendustes, nt teksti analüüs ja süntees, otsisüsteemid, õigekeelsuse kontroll, keeleõpe, automaattõlge, lingvisti töövahendid jne.

Nende nõuete täitmise teevad võimalikuks 3 põhiprintsiipi, mida avatud morfoloogiamudeli arendamisel on järgitud:

- reeglipõhisus,
- andmete ja tarkvara sõltumatus,
- modulaarsus.

Reeglipõhisus loob võimaluse luua keeleliselt usaldusväärne ja avatud süsteem. Süsteem on üles ehitatud nii, et igale lingvistilisele allsüsteemile vastab oma reeglite interpretaator ning omad reeglid ja erandiloend(id). Iga moodulit saab arendada eraldi, võttes aluseks vastavas valdkonnas tehtud lingvistilised uuringud. Erandite loendi moodustamine toimub paralleelselt reeglite testimisega. Erandite hulk keeles on suhteliselt väike ja kindlapiiriline – need on võimalik esitada lõplike loenditena. Valdav osa keelenähtusi on kirjeldatavad reeglitega, ja eriti käib see uuemate sõnade ja sõnavormide kohta: tundmatu sõna käitub suure tõenäosusega reeglipäraselt.

Avatud süsteem on eesti keele puhul praegu eriti aktuaalne, sest suured muutused ühiskonnas mõjutavad paratamatult ka keelt. Osa nähtusi kaob – kaovad ja ununevad ka sõnad, osa nähtusi tekib – tekivad ka uued sõnad. Uuenenud on ka kontaktid teiste keeltega: vene keele mõju on taandunud ja asendunud inglise (ja kohati ka soome) keele pealetungiga. Aja jooksul hakkab see ka keele grammatilist süsteemi mõjutama, nii et on vaja muuta reegleid.

Andmete ja tarkvara sõltumatus tagab süsteemi paindlikkuse ja avatuse. Tarkvara põhikomponent on reeglite interpretaator, mille jaoks on oluline ainult õige reegliformaat, mitte aga reeglite hulk või sisu. Kõik andmed (reeglid, juhtinfo ja erandisõnastikud) esitatakse tavalise tekstina, mida on võimalik täiendada ja korrigeerida nii, et süsteemi töökorraldus sellest ei muutu.

Nii sõnastikke kui ka reegleid saab kohandada vastavalt sellele allkeelele, mille töötlus parajasti käsil on – olgu selleks siis teaduslik artikkel geneetikast, tolliseaduse tekst või ulmelugu Maavälistest olenditest. Seejuures jääb tarkvara ikka samaks ning olemasolevatele moodulitele saab vajaduse korral lisada uusi. Sama tarkvarasüsteem oleks võimalik panna tööle ka teiste keelte jaoks, kui asendada kõik sõnastikud ja reeglid ning modifitseerida juhtinfot.

Modulaarsus tagab süsteemi paindlikkuse eri rakenduste suhtes ning ühtlasi võimaldab kontrollida süsteemi adekvaatsust. Iga lingvistiline allsüsteem vormistatakse iseseisva programmimoodulina (nn dünaamilise teegina `.dll`), mida saab teistest programmidest välja kutsuda. Mooduleid saab omavahel erinevalt kombineerida, või ka ükshaaval kasutada – vastavalt konkreetse rakenduse vajadustele. Moodulitel on mitu valitavat töörežiimi, nt tüübituvastus ja tüvemuutused töötavad kas vältega või ilma välteta režiimis (välte arvestamine on oluline nt kõnesünteesis, keeleõppes ja nõudlikumas leksikograafias).

Sellises keerulises modulaarses süsteemis sõltub ühe mooduli väljundi kvaliteedist iga järgmise mooduli töö. Kui süsteemi lõppväljund on pärast mitme mooduli läbimist siiski ootuspärane ja lingvistiliselt korrektne, siis võib oletada, et loodud mudel on vähemalt mingis mõttes adekvaatne tegeliku keelesüsteemiga.

3.3. Tulemused

Töö avatud morfoloogiasüsteemiga algas 1993. a Avatud Eesti Fondi toetusel, hiljem on projekti rahastanud Eesti Teadusfond ja osaliselt keeletehnoloogia sihtprogramm. Süsteem on tegemise käigus koos tegijatega arenenud ja muutunud ning nii mõnigi moodul on saanud uue kuju. Nüüdseks on süsteemi olulisemad moodulid realiseeritud dünaamiliste teekidena ning on EKI koduleheküljele vabaks kasutamiseks välja pandud. Moodulite peamised autorid on Indrek Hein ja Evelin Kuusik.

Morfoloogiamoodulid on EKI serveril kahe paketina:

- http://www.eki.ee/tarkvara/:
  - syllabif silbitus
  - typedet tüübi ja sõnaliigi tuvastus
  - stems tüvemuutused, sh lemmatiseerimine
  - fmsynth morfoloogiline vormisüntees
- http://www.eki.ee/keeletehnoloogia/projektid/morfana/:
  - ana morfoloogiline analüüs
  - LS-piir liitsõnapiiride tuvastus

Iga mooduliga on näidiseks kaasas ka üks lihtne tarbijaprogramm.

Morfoloogiamooduleid on seni kasutatud peamiselt lingvistilises uurimistöös, nt tüvemuutuste ja tüübituvastuse süsteemide modelleerimisel, silpide andmebaasi loomisel, grammatilise homonüümia uurimisel, sõnavormide kasutuse uurimisel jne. Olulisematest praktilistest rakendustest võiks nimetada grammatilist kirjegeneraatorit, mille abil saab sõnastiku sõnaartiklitesse poolautomaatselt lisada grammatilised andmed (sõnaliik, muuttüüp, muutevormid, üksuste piirid, grammatilised viited jne) eesti märksõnade või vastete jaoks (Kuusik, Lind, Viks 1995; Viks 2000).

Kirjandus

Hein, I. 1994. Practical realisation of the morphological analysis. – Automatic Morphology of Estonian 1. (Research Reports.) Toim Ü. Viks. Tallinn: Eesti Keele Instituut. 29–35.

Hein, I. 1995. Rules for finding boundaries in compound words. – Automatic Morphology of Estonian 2. (Research Reports.) Toim Ü. Viks. Tallinn: Eesti Keele Instituut. 7–22.

Kaalep, H.-J. 1996. ESTMORF: A morphological analyzer for Estonian. – Estonian in the Changing World. Toim. H. Õim. Tartu: Tartu Ülikooli Kirjastus. 43–98.

Kuusik, E. 1994. Morphological synthesis of Estonian based on the agglutination strategy. – Automatic Morphology of Estonian 1. (Research Reports.) Toim Ü. Viks. Tallinn: Eesti Keele Instituut. 36–48.

Kuusik, E. 1995. Automatic recognition of the Estonian stem changes. – Automatic Morphology of Estonian 2. (Research Reports.) Toim Ü. Viks. Tallinn: Eesti Keele Instituut. 46–71.

Kuusik, E. 1996. Eesti tüvemuutuste süsteemi modelleerimine. Magistriväitekiri (käsikiri Eesti Keele Instituudis).

Kuusik, E., Lind, P., Viks, Ü. 1995. An Estonian Morpho-Generator for Dictionaries. (Preprint FU 1995.) Tallinn: Eesti Keele Instituut.

Kuusik, E., Viks, Ü. 1998. Reeglipõhine morfoloogiline süntees. – Arvutimaailm 1, 43–45, 63; 2, 19–21.

Viks, Ü. 1977. Klassifikatoorse morfoloogia põhimõtted. (Preprint KKI-9.) Tallinn: Eesti Keele Instituut.

Viks, Ü. 1978. Morfoloogilise klassifikatsiooni optimeerimisest. – Sõnast tekstini. Tallinn: Eesti Keele Instituut. 91–111.

Viks, Ü. 1992. Väike vormisõnastik I: Sissejuhatus & grammatika; Väike vormisõnastik II: Sõnastik & lisad. Tallinn: Eesti Keele Instituut.

Viks, Ü. 1994. Eesti keele morfoloogiline analüsaator. Automaatanalüüsi võimalused ja võimatused. – Keel ja Kirjandus 3, 150–163.

Viks, Ü. 1995a. About rule-oriented morphology of Estonian. – Abstracts of Posters Presented at the 10th Nordic Conference of Computational Linguistics NODALIDA–95. Helsinki. 28–30.

Viks, Ü. 1995b. Rules for recognition of inflection types. – Automatic Morphology of Estonian 2. (Research Reports.) Toim Ü. Viks. Tallinn: Eesti Keele Instituut. 23–45.

Viks, Ü. 1997. Erand, reegel ja sõnastik avatud morfoloogiamudelis. – Pühendusteos Huno Rätsepale. (Tartu Ülikooli eesti keele õppetooli toimetised 7.) Toim. M. Erelt, M. Sedrik, E. Uuspõld. Tartu: Tartu Ülikooli Kirjastus. 244–254.

Viks, Ü. 2000. Tools for the generation of morphological entries in dictionaries. – Proceedings of the 2nd International Conference on Language Resources and Evaluation LREC2000. Athens.

Kahetasemeline morfoloogiamudel eesti keele arvutimorfoloogia alusena

Heli Uibo
Tartu Ülikool

1. Sissejuhatus

Loomuliku keele **morfoloogilise analüüsi** sisendiks on sõnavorm, väljundiks algvorm ehk lemma ja grammatilised tähendused. **Morfoloogilise sünteesi** sisendiks on algvorm ja grammatilised tähendused ning väljundiks sõnavorm.

Järgnevalt anname loetelu morfoloogilise analüüsi ja sünteesi kasutusvõimalustest, mida võiks veelgi jätkata (Sproat 1992: 1–14):

- täielikku loomuliku keele töötlust eeldavad rakendused: keeleanalüsaatorid ja -generaatorid, masintõlge (viimaseks on vajalikud nii lähtekeele analüüs kui resultaatkeele süntees);
- sõnastike koostamise vahendid (õigekeelsus- või vormisõnastikud, erialaterminoloogiasõnastikud jne.);
- kõnesünteesi- ja kõnetuvastusvahendid;
- õigekirjakontrollijad;
- infootsisüsteemid (Seal on lemmatiseerimine väga vajalik, kuna algvormis esitatud päring on efektiivsem, konkreetse morfoloogilise sõnavormi otsing omab tavaliselt vähe mõtet. Ilma morfoloogiatöötluseta infootsisüsteemides tuleb päringud anda ette kujul "tüvi*" so. jätta muutuv tüvelõpp ja formatiivid ära, aga eesti keeles muutuvad astmevaheldusliкel sõnadel ka tüve sisehäälikud.);
- lingvistilise uurimistöö vahendid (nt keelekorpuse märgendamiseks ja uurimiseks, tekstide sõnavara analüüsiks, mis eeldab lemmatiseerimist jne);
- keeleõppeprogrammid – sh nii keele kui suhtlemisvahendi omandamiseks kui ka filoloogiatudengitele morfoloogiakursuse tarvis;
- jne.

Eesti keele morfoloogia automatiseerimisega on tegeldud ja tegeldakse nii Eesti Keele Instituudis (EKI) kui Tartu Ülikoolis (TÜ). EKIs on arvutimorfoloogia-alane juhtivteadur Ülle Viks, kes on loo-

nud uue, kujundite tuvastamise teoorial põhineva morfoloogilise klassifikatsiooni. Seda käsitleb doktoriväitekiri "Klassifikatoorne morfoloogia" (Viks 1994a), mis ühtlasi võtab kokku eelneva töö morfoloogilise klassifikatsiooni loomisel – verbide ja noomenite morfoloogilise klassifikatsiooni ning "Väikese vormisõnastiku" (Viks 1980, 1982, 1992).

Uue lähenemise eesti keele automaatsesse morfoloogiasse tõid Evelin Kuusiku tüübituvastusreeglid (Kuusik 1996; Kuusik, Viks 1998). See tähendab **avatud morfoloogiamudeli** idee realiseerimist: kõik produktiivsed ja regulaarsed morfoloogianähtused kirjeldatakse ja realiseeritakse aktiivse morfoloogia reeglite abil, sõnastikus esitatakse vaid erandid. Tundmatud sõnad püütakse analüüsida tüübituvastusreeglite abil ja enamasti see ka õnnestub, kuna uued keelde tulevad sõnad muutuvad reeglipäraselt.

Heiki-Jaan Kaalepi morfoloogilise analüüsi programm ESTMORF (Kaalep 1999) ning MS Office õigekirjakontrollija põhinevad samuti suures osas "Väikesel vormisõnastikul"

Nagu eelnevast selgub, on eesti keele morfoloogia põhiliste rakenduste jaoks juba formaliseeritud. Sellegipoolest polnud seniajani praktilist kogemust, kas ja kuidas võimaldab eesti keele morfoloogiat kirjeldada **kahetasemeline morfoloogiamudel**, mis paljude keelte puhul on osutunud edukaks. Seda lünka püüab käesolev uurimus täita. Teemat on esmakordselt käsitletud artiklis (Uibo 1998), kus on antud ülevaade kahetasemelise morfoloogia formalismist, mõningane eesti keele morfoloogia iseärasuste analüüs ja hüpoteesid selle kohta, milliseid eesti keele morfoloogia nähtusi võiks kirjeldada reeglitega ja milliseid sõnastikega. Vahepeal on uurimistöös toimunud märgatav areng – on loodud reeglitest ja sõnastikest koosnev eksperimentaalne eesti keele morfoloogia kirjeldus, mille kooskõlalisus on saavutatud praktilise testimise teel.

2. Ülevaade kahetasemelisest morfoloogiamudelist

2.1. Mudeli põhiolemusest

Kahetasemelise morfoloogiamudeli esitas 1983. aastal oma doktoriväitekirjas praegune Helsingi Ülikooli arvutuslingvistika professor Kimmo Koskenniemi (Koskenniemi 1983).

Klassikaline **generatiivne fonoloogia** põhineb ülekirjutusreeglitel. Ülekirjutusreeglid on kujul **uXv** → **uYv**. See tähendab, et mitte-

terminaalne sümbol **X** asendatakse sümboliga **Y**, kui sümboli **X** vasakpoolseks kontekstiks on **u** ja parempoolseks kontekstiks **v**. Ülekirjutusreeglitest koosnev grammatika peab reeglite järjest rakendamisel genereerima keele kõik lubatavad sõnavormid, ja ainult need.

Ka kahetasemeline morfoloogiamudel kasutab teisendusreegleid, kuid reeglite rakendamise järjekord pole oluline, sest iga reegel töötab ülejäänutest sõltumatult.

Generatiivse fonoloogia üheks puuduseks on ka see, et tuletusprotsessi ajal kättesaadavad fonoloogilised tunnused ei suuda kirjeldada kogu selle taga olevat morfoloogiat (Karlsson 1974). Seevastu kahetasemelistele reeglitele on kättesaadavad nii fonoloogilised kui morfoloogilised tunnused, mis on tegelikult mõlemad aluseks sõnade klassifitseerimisel muuttüüpidesse.

Mudeli muudab kahetasemeliseks asjaolu, et sõnastikus ei säilitata mitte sõnavormide koostisosi, vaid nende nn **süvaesitusi**, millest saab sõnastikevahelisi viitade ja kahetasemeliste reeglite abil moodustada tegelikud kirjakeelsed sõnavormid (ehk sõnavormide nn **pindesitused**). Vaatleme sõnavormi *mõtetes* süva- ja pindesitust:

süvaesitus:	m õ t T e + t e + s #
pindesitus:	m õ t 0 e 0 t e 0 s 0

Kokkuleppeliselt märgitakse süvaesituses suurtähtedega morfofoneeme, millel on pindesituses rohkem kui üks variant. Nii on siin *T*-ga tähistatud süvaesituse neljas foneem, mis tugeva astme vormides vastab *t*-le, nõrga astme vormides aga kaob. Morfofoneeme, millel on ainult üks variant, märgitakse ka süvaesituses väiketähega. Süvaesitusse saab märkida ka morfoloogilisi tunnuseid ja morfeemide ning liitsõnaosade piire. Märk "+" viitab käändelõpu või tunnuse järgnemisele ning märk "#" tähistab sõnavormi lõppu. Seda tüüpi sümbolitele seatakse pindesituses vastavusse tühisümbol "0", et süva- ja pindesitus rajastada.

Märkimata ei saa jätta ka kahetasemelise morfoloogia **kahesuunalisust** (Koskenniemi 1983) – mudel kujutab endast keele morfoloogilise süsteemi kirjeldust ning pole otseselt orienteeritud morfoloogilisele analüüsile ega sünteesile. Seejuures nii analüüsi kui sünteesi realiseerimiseks on olemas efektiivne algoritm.

Ü. Viksi artiklis (Viks 1994b) käsitletud automaatse morfoloogilise analüüsi strateegiatest kasutab kahetasemeline mudel teisen-

dusstrateegiat, mis sisaldab endas ühtlasi otsingu- ja tuletus-liigendusstrateegiat:

1) otsing – tüvede sõnastikest otsitakse sõnavormi algusotsaga kokkulangevat sõnet;
2) tuletus-liigendus – sõnavormi tuletamiseks liigutakse sõnastike võrgus mööda morfotaktika reeglitele vastavaid viitu;
3) teisendus – kahetasemelised reeglid teisendavad pindesituse süvaesituseks ja vastupidi.

Kahjuks ei ole kahetasemelises mudelis kohta tuvastusreeglitele, mistõttu mudeli baasil realiseeritud süsteemi sõnastik vajab aeg-ajalt uuendamist. Lahenduseks oleks süsteemiga kaasaskäiv sõnastiku uuendamisprogramm, mis teeks kindlaks tundmatute sõnade tüübid ning koostaks korrektsed tüvedesõnastiku kirjed. Teisendusstrateegia kasuks võrreldes tuletus-liigendusstrateegiaga räägib eesti keele puhul asjaolu, et eesti keeles on mitut liiki tüvemuutused väga levinud ning suurem osa sõnadest kahe- (*luge|da-loe|n, kallas-kalda*) või kolmetüvelised (*käski|da-käsi|n-käs|takse, sõber-sõbra-sõpra).* Pelgalt tuletus-liigendusstrateegiat kasutades peaksid sõnastikus leiduma kõik erinevad tüvevariandid, seega tüvedesõnastiku kirjeid oleks umbes kaks korda rohkem kui vastavaid algvorme.

Nüüdseks on kahetasemeline mudel maailmas üldtuntud ning leidnud rakendust paljude keelte (inglise, saksa, rootsi, soome, suahiili jt) morfoloogia automatiseerimisel. Mudelil põhineva keeletarkvara väljatöötamisega tegelevad Soome Helsingi Ülikooli Üldkeeleteaduse osakond ja Soome keeletehnoloogiafirma Lingsoft (http://www.lingsoft.fi/) ning Xeroxi Euroopa Uurimiskeskus Grenoble'is (http://www.xrce.xerox.com/research/mltt/). Firmas Lingsoft on katsetusi tehtud ka eesti keelega, kuid ilma reegliteta: kogu kirjeldus põhineb sõnastikel.

2.2. Kahetasemelised reeglid

Kahetasemelised reeglid peavad ära kirjeldama kõik erinevused süva- ja pindesituse vahel. Tervet reeglite kogu võib vaadelda kui filtrit, läbi mille näeb sõnavormi süvaesitust pindesitusena ja vastupidi. Reeglite rakendamise järjekord pole määratud: kõik keelekirjeldusreeglid peavad olema rahuldatud üheaegselt. Reeglid realiseeritakse sümbolipaaride jadasid töötlevate lõplike teisendajatena.

Lõplikud automaadid ja lõplikud teisendajad on loomuliku keele töötluses osutunud väga efektiivseteks: nende abil on realiseeritud nii sõnastikke, morfoloogilise analüüsi, morfoloogilise ühestamise kui süntaktilise analüüsi reegleid (Roche, Schabes 1997).

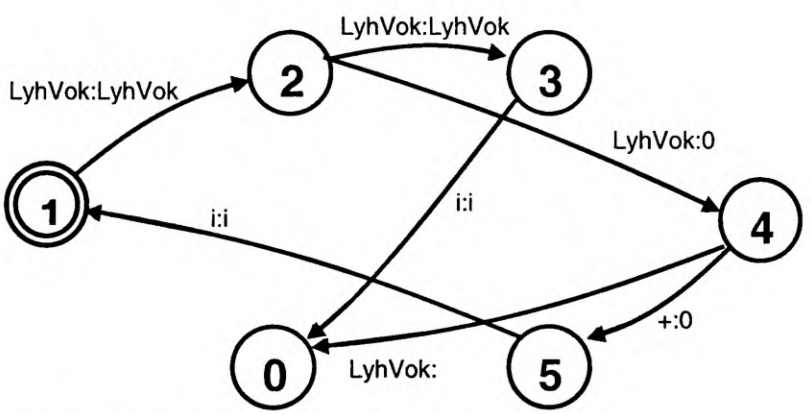
Lõplikku teisendajat võib tõlgendada lõpliku automaadina, mille sisendiks on sümboli**paaride** (mitte üksiksümbolite) järjend ehk mille iga kaar on märgendatud sümbolipaariga, mitte ühe sümboliga. Seega kahetasemeliste reegel-automaatide sisendiks on sümbolipaaride järjendid, kus paari esimene komponent on süvaesitusest ning teine komponent pindesitusest.

Kleene'i teoreemi kohaselt (Roche, Schabes 1997: 15) on regulaarsed avaldised on samaväärsed lõplike automaatidega.

Lõplikku teisendajat on võimalik kirja panna:

a) olekugraafina (definitsioon olekuid ühendavate kaarte kaudu);
b) olekuteisendustabelina (definitsioon olekuteisendusfunktsiooni kaudu);
c) regulaarse avaldisena (Kleene'i teoreem).

Lõpliku teisendaja töö illustreerimiseks esitame reegli "Pikk vokaal lüheneb enne i-ga algavat formatiivi" (nt *maa-maid, tee-teid*) kõigil kolmel viisil. Suvalise *i*-st erineva täishääliku märkimiseks defineerime potentsiaalsete lühenevate vokaalide hulga LyhVok = {a, e, o, u, õ, ä, ö, ü}.



Joonis 1. Lõpliku teisendaja esitus olekugraafina

Graafis on ülevaatlikkuse huvides märkimata jäetud kõik silmused ehk kaared, mis algavad ja lõpevad samas olekus. Kui sisendist

loetav sümbolipaar ühtib mingi antud olekust lähtuva kaare märgendiga, siis läheb automaat uude olekusse, milles see kaar lõpeb. Kõikvõimalike muude sisendpaaride korral jääb automaat endisesse olekusse.

Oleku nr.	+ 0	i i	LyhVok LyhVok	LyhVok 0	= = **
1.	1	1	2	1	1
2:	2	2	3	4	1
3:	3	0	3	3	3
4:	5	4	0	0	4
5:	5	1	5	5	5

Joonis 2. Lõpliku teisendaja esitus olekuteisendustabelina
* Lõppoleku numbri järel on punkt, teiste järel koolon.
** Võrdusmärgid tähistavad kõikvõimalikke ülejäänud sümbolipaare, mis ei oma antud reegli seisukohast tähtsust.

Regulaarne avaldis on kolmest esitusest kõige hõlpsamini kirjapandav ja loetav: V1:0 ⇔ V1 _ %+: i ; where V1 in LyhVok;

Reeglit tuleb lugeda nii: "Hulgas LyhVok sisalduvale täishäälikule vastab pindesituses 0 (ehk ta kaob) parajasti siis, kui tema vasakpoolseks kontekstiks on seesama täishäälik ning vahetu parempoolse konteksti süvaesituses + ja seejärel *i* nii süva- kui pindesituses"

Mudeli praktilisel realisatsioonil XRCE tarkvara abil kasutataksegi kahetasemeliste reeglite esitust regulaaravaldistena.

Kahetasemelise reegli üldkuju (Koskenniemi 1983) on

CP op LC _ RC, kus op ∈ {⇐, ⇒, ⇔}

Seejuures CP[1] on sümbolipaar, LC – vasakpoolne kontekst ja RC – parempoolne kontekst.

Nii reegli vasakul kui paremal poolel võib esineda loetelu. Vasakul pool võib olla mitu sümbolipaari:

$CP_1, CP_2, ..., CP_n$ op LC _ RC

Seda interpreteeritakse kui n reeglit, kus vastavused on erinevad, kuid kontekst sama.

[1] Paneme tähele, et CP, LC ja RC on üldjuhul hulgad, mitte konkreetsed sümbolid või sümbolijärjendid. Nii on lk 41 kasutatud hulka LyhVok = {a, e, o, u, õ, ä, ö, ü) ja lk 43 hulki Vok = {a, e, i, o, u, õ, ä, ö, ü}, Kons = {b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, x}, TyveVok = {a, e, i, u} ja Liq = {l, r}.

Üks ja sama süva- ja pindsümboli vastavus võib esineda erinevates kontekstides. Seegi on kirjapandav ühe reegliga, mille paremal poolel on loetletud kõik võimalikud kontekstid:

CP op LC_1 _ RC_1
LC_2 _ RC_2
...
LC_n _ RC_n

Näide kontekstide loendit sisaldavast reeglist "D kadu nõrgas astmes" on järgmine:

```
D:0   ⇔   Algus Vok Vok: _ (TyveVok:) %$:;      !laud-laua
          Vok Vok Liq _ TyveVok %$:;            !siirdama-siirata
[ e | i: | u: | ü: ] _ TyveVok: %$: ;           !vedama-vean, rida-rea
                      õ _ e %$: ;               !õde-õe, põdeda-põen
                      [Kons [r|s] ] a _ u: %$: ;  !kaduda-kaon
```

Reegel "D kadu nõrgas astmes" tähendab, et sümbolile D ei vasta pindesituses midagi siis ja ainult siis, kui ta esineb ühes loetletud kontekstidest. Allkriips _ märgib paari D:0 kohta vasak- ja parempoolse konteksti vahel. Et muuta reegleid paremini loetavaiks, on soovitav kasutada sümbolite hulki, näiteks on defineeritud kõikide eesti keeles kasutatavate täishäälikute hulk **Vok**, võimalike tüvevokaalide hulk **TyveVok** ja konsonantide hulk **Kons**. Ka enamkasutatavatele sõnasegmentidele saab anda nimesid, näiteks **Algus** tähistab silbi, sõna või liitsõnakomponendi algust. Paneme tähele, et kõik kontekstid lõpevad sümboliga **$**„ st D kaob ainult astmevahelduslike sõnade nõrgas astmes. Hüüumärgi järele on kirjutatud konkreetne näide iga konteksti kohta, mis hõlbustab reeglist arusaamist.

2.2. Sõnastikud

Sõnastike võrk defineerib keele morfotaktika- ja allotaktikareeglid. Sisu järgi võib eristada järgmisi sõnastikke:
1) tüvikusõnastikud;
2) tüvelõpumuutuste sõnastikud;
3) tunnuste ja lõppude sõnastikud;
4) hargnemissõnastikud.

Seejuures kõikide sõnastike kirjestruktuur on ühesugune. Esialgu, Koskenniemi doktoritöös (Koskenniemi 1983: 28) oli välja pakutud veidi erinev sõnastiku struktuur. Muudatus tuli sellest, et sõnastikud on nüüd realiseeritud kui **sõnastik-teisendajad** *(lexical transducer)*.

Sõnastik-teisendaja toetab nii morfoloogilist analüüsi kui sünteesi. Morfoloogiline analüüs toimub järgmise algoritmi järgi (Sproat 1992: 6).

Algoritm 1. Morfoloogiline analüüs sõnastik-teisendajas. Sisend: analüüsitava sõnavormi süvaesitus, Väljund: algvormi süvaesitus ja morfoloogiline info või veateade.

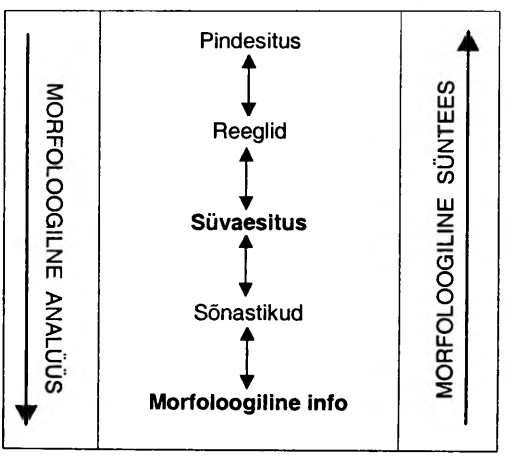
Samm 1. Sõnastik-teisendaja alustab algolekust q_1.

Sammu 2 tehakse tsükliliselt, kuni teisendaja on kaari mööda liikudes jõudnud sõne lõppu (leidus edukas tee) või sõnet ei suudetud sõnastike võrgus liikudes genereerida (tulemus: sõne on tundmatu või vigane).

Samm 2. Kui sisendsõne algusosa ühtib mõne kaare märgendi alumise poolega, liigub teisendaja vastavat kaart mööda järgmisse olekusse, seejärel lõikab sisendsõne algusest ära juba leitud jupi ja väljastab selle jupi kohta käiva grammatilise informatsiooni.

Morfoloogilise sünteesi puhul on sisendiks algvormi süvaesitus ja grammatiline info ning sobiva sisendi korral väljastab ta juppjupilt vastava sõnavormi süvaesituse.

Selleks, et morfoloogilise analüüsi ja sünteesi sisendiks ja väljundiks oleks pindesitus, tuleb sõnastik-teisendaja ja reeglid ühendada (joonis 3).



Joonis 3. Morfoloogiline analüüs ja süntees kahetasemelises mudelis

Sõnastik-teisendaja on kompileeritav tekstifailist, milles iga sõnastiku struktuur peab olema järgmine:

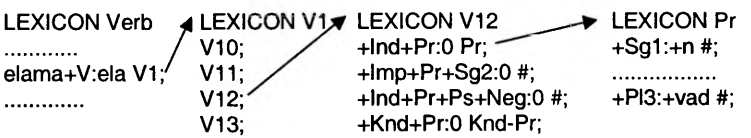
```
<sõnastik> ::= "LEXICON" <nimi>
                <kirjete loetelu>
<kirjete loetelu> ::= kirje [<kirjete loetelu>]
kirje ::= <morfoloogiline info> ":" <süvaesitus> <jätkuviit> ";"
```

Jätkuviit näitab, millisesse sõnastikku on antud punktist võimalik edasi liikuda. Jätkuviitade abil ühendatud sõnastike võrk võimaldab arvestada morfeemide sõnasiseseid järgnevusi.

```
Kindla kõneviisi oleviku pöördelõppude sõnastik eesti keeles
LEXICON Pr
+Sg1:+n #;
+Sg2:+d #;
+Sg3:+b #;
+Pl1:+me #;
+Pl2:+te #;
+Pl3:+vad #;
```

Märk "#" jätkuviida kohal tähendab, et midagi rohkemat sõnavormile liituda ei saa. Sõnastike süsteemis on olemas ka sõnastikud, mis viitavad sõnastikule Pr – nendeks on erinevate pöördkondade kindla kõneviisi oleviku tüvest moodustatavate vormide sõnastikud, nt V12 (joonis 4). Vastava morfoloogilise märgendi +Ind+Pr saavad sõnad just sealt külge. Algvorm ja sõnaliik (+V = verb) aga saadakse kätte verbitüvede sõnastikust, kus tüve sõnastikuesituse järel on viit esimese pöördkonna jätkusõnastikule V1.

Sõnastikevahelisi seoseid illustreerib joonis 4.



Joonis 4. Kindla kõneviisi oleviku moodustamine verbist *elama* sõnastikevahelisi viitu järgides

3. Eesti keele morfoloogia kirjeldamine kahetasemelise morfoloogiamudeli vahenditega

3.1. Eesti keele morfoloogia põhijooni

Eesti keel on tugevalt **morfoloogiline keel** (Kuusik, Viks 1998) – grammatilisi tähendusi väljendatakse tunnuste ja lõppude abil, mis liituvad tüvele (mitte ees- ega tagasõnade abil).

Eesti keelele on suures osas omane aglutinatiivsus – sõnavormid moodustuvad morfoloogiliste formatiivide liitumisel sõnatüvele. Sõnavorm liigendub tüveks, mis väljendab leksikaalset tähendust ja formatiiviks, mis väljendab grammatilist tähendust. Detailsema liigenduse kohaselt koosneb tüvi juurest ja tuletusliidetest ning formatiiv üksikmorfeemidest (tunnustest ja lõppudest).

Sõnavormi *juhuslikkudest* liigendamine

Juhuslikku +	dest	juhus +	likku	+de	+st
Tüvi +	formatiiv	juur +	tuletusliide	+tunnus	+käändelõpp

Eesti keeles võib eristada järgmisi morfoloogilisi sõnaklasse:

- käändsõnad ehk noomenid,
- pöördsõnad ehk verbid,
- muutumatud sõnad ehk indeklinaablid.

Noomenitel on (erinevate teooriate kohaselt) 14–15 käänet ainsuses ja mitmuses, sageli mitmuses ka paralleelvormid *i*-mitmuse või tüvemitmuse näol. Ü. Viksi eeskujul on lühike sisseütlev kirjeldatud iseseisva käändena – suunduva ehk aditiivina, mis pole küll kõikide sõnatüüpide puhul moodustatav.

Verbidel on neli kõneviisi (kindel, tingiv, käskiv, kaudne), neli aega (olevik, lihtminevik, täisminevik ja enneminevik), kaks tegumoodi (isikuline ja umbisikuline), kaks kõneliiki (jaatav ja eitav), kolm isikut (1., 2., 3.) ning kaks arvu (ainsus ja mitmus).

Eesti keeles liigitatakse sõnad morfoloogilistesse muuttüüpidesse enamasti järgmiste tunnuste alusel.

1. Astmevaheldus (tüve sisehäälikute ja/või välte muutumine): nõrgenev, tugevnev, puudub.

Astmevahelduseta sõnadel võib muutuda ainult tüvelõpp (*raamat-raamatu, kõnelda-kõnelen*) või jääb tüvi täiesti muutumatuks (*karu-karu, kirjutada-kirjutan*).

Astmevahelduse võib täpsemalt liigitada

a) laadivahelduseks (häälikukadu (*nägu-näo*), assimilatsioon (*lendama-lennata*) või asendumine (*sadama-sajab*) nõrgas astmes);

b) klusiilide (k, p, t) vältevahelduseks (*pilt-pildi*);

c) prosoodiliseks vältevahelduseks (*vein* (III v) – *veini* (II v)).

2. Tüvelõpumuutused (Viks 1979):

a) lisandumine või kadu: tüvevokaal (*klass-klassi*), tüvelõpu-*S* (*vilgas-vilka*), silp *-da* või *-me* (*tore-toreda, mõõde-mõõtme*);

b) hääliku(ühendi) asendumine: *ne-se* (*inimene-inimese*), *ne-sa* (*õnnis-õndsa*), *s-ne* (*küüs-küüne*), *s-kse* (*juus-juukse*), *i-e* (*lumi-lume*), *le-el* (*vaidlema-vaielda*) jt.;

c) nii kadu kui lisandumine: *number-numbri*, *armas-armsa*

3. Silpide arv: 1, 2, 3 ja enam.

4. Välde: I, II, III.

5. Morfoloogiliste formatiivide valik (nt mitmuse tunnus *-de* või *-te*, umbisikulise tegumoe tunnus *-t* või *-d*).

6. Erinevate tüvekujude paiknemine paradigmas (tugeva- ja nõrgaastmelise tüve ning lemma- ja muutetüvede paiknemine paradigmas – vt näidet jaotisest 3.3).

3.2. Kahetasemeliste reeglite koostamine eesti keele jaoks

Kuna iga reegel vaatleb korraga vaid ühte süva- ja pindsümboli vastavust, on loomulik kirjeldada reeglitega ühte häälikut puudutavad häälikumuutused. Kahetasemelised reeglid ei suuda korraga teisendada ühte tervet segmenti teiseks ning teisenduse läbiviimine mitme reegli abil nõuab reeglite omavahelist koordinatsiooni. Seepärast kirjeldatakse "ebaloomulikud" fonoloogiliselt ja morfoloogiliselt põhjendamatud häälikumuutused minisõnastikega. Reeglitega saab kirjeldada nii fonoloogiast kui morfoloogiast tingitud häälikumuutusi, kuna kontekstides võivad esineda nii pind- kui süvaesituse sümbolid. Fonotaktilise häälikumuutuse korral piisab enamasti pindkontekstist, morfo(fono)loogiline muutus aga nõuab ka süvaesituse kaasamist konteksti.

Eesti keele puhul osutub kasulikuks see, et reeglite kontekstides saab kasutada sõnaosade piire märkivaid tunnuseid. Näiteks astmevahelduse reeglid vaatlevad ainult sõnatüve sisehäälikuid ning tüve

lõppu paigutatavat nõrga astme tunnust, suur osa reeglitest käivitub just tüve ja tunnuse/lõpu piiril või liitsõnapiiril (nt reegel: "umbisikulise tegumoe tunnus d → t pärast s-i või h-ga lõppevat tüve").

Järgnevalt kommenteerime koostatud eesti keele kahetasemelisi reegleid põhjalikumalt.

Reeglitefaili alguses kirjeldatakse kasutatav tähestik, kus on üles loetud esiteks kõik pindesituse sümbolid ning seejärel võimalikud paarid, millesse süvaesituse sümbolid võivad kuuluda. Tüüpiliselt on üks sama süvaesituse sümbolit sisaldavatest paaridest nn vaikepaar (*default pair*), teis(t)e esinemist piiravad reeglid.

Eesti keele kahetasemeliste reeglite kogu sõnastik on järgmine:

a b c d e f g h i j k l m n o p q r s sh z zh t u v õ ä ö ü x y G:g G:0 B:b B:0 T:d T:0 S:s S:0 S:r %+:0 %$:0 %&:0 A:a A:0 E:e E:0 U:u U:0 %.:a %=:0 2:0

Edasi defineeritakse vajalikud tähestiku alamhulgad, näiteks juba eespool esinenud Kons, Vok, TyveVok, LyhVok, lisaks astmevaheldust markeerivate suurtähtede hulk AV jm.

Kui mingi alamavaldis esineb reeglites sageli, võib talle anda nime, eriti kui see alamavaldis omab sisulist tähendust. Nii on kirjeldatud morfeemipiir **Piir** = [.#. | %+: | %&:] ja sõna, liitsõna komponendi või silbi algus: **Algus** = [.#. | Kons+ | %&:]

3.2.1. Astmevahelduse reeglid

Tänapäeva eesti keeles ei määra sõna häälikuline kuju enam tema astmevaheldusliikkust (Hint 1978). Astmevaheldus on tänapäeva eesti keeles leksikaalselt tingitud, ta on sõnade eriomadus, mis ei tulene paratamatult sõna häälikulisest struktuurist. Seega tuleb astmevaheldusliikud sõnad märgendada, seejuures ära märkides ka selle, missugused vormid on tugevas ja missugused nõrgas astmes.

Kahetasemeline morfoloogiamudel annab võimaluse astmevaheldust siiski reeglitega käsitleda. Kuna reegel vaatleb korraga sümboli**paari**, mis koosneb vastavatest süva- ja pindesituse sümbolitest, saab reeglis viidata ka muudele sümbolitele peale otseselt sõnavormis esinevate häälikute. Astmevaheldusliike sõnade puhul märgitakse muutuv foneem süvaesituses mingi pindtähestikus mitte sisalduva sümboliga, võimaluse korral tugevale astmele vastava suurtähega. Sellel suurtähel on pindesituses kaks varianti, mis vastavad parajasti nõrgale ja tugevale astmele. Tegemaks valikut nõrga ja tugeva astme vahel, markeeritakse nõrgaastmeline tüvi sümboliga

"$" Astmevahelduse reegel, kohates sobivat suurtähte, otsib sõnatüve lõpust nõrga astme markerit. Kui marker leidub, aktsepteeritakse reeglis esinev sümbolipaar, vastasel korral läheb käiku reeglitekogu tähestikus vaikimisi kehtestatud vastavus.

Nõrgaastmelise tüve markeri "$" mõju

k o o K i $ + d	ja	k o o K i + d e + s t	aktsepteeritakse
k o o g i 0 0 d		k o o k i 0 d e 0 s t	
k o o K i $ + d	ja	k o o K i + d e + s t	ei aktsepteerita
k o o k i 0 0 d		k o o g i 0 d e 0 s t	

Sümboliga "$" markeeritud tüves sobib nõrgale astmele vastav paar K:g, markeerimata tüves läheb läbi vaikimisi kehtiv vastavus K:k.

Esmapilgul näib, et astmevahelduse reeglid võivad siis olla kõik ühesugused. Reegli vasakul poolel on vastavus <suurtäht>:<nõrgale astmele vastav häälik>, vasak kontekst pole oluline ning kusagil paremas kontekstis peab leiduma marker "$" Kuid eesti keele astmevaheldussüsteemis pole vastavus tugeva ja nõrga astme allomorfide vahel üksühene: ühele ja samale tugeva astme allomorfile võib nõrgas astmes vastata erinevaid häälikuid.

Astmevaheldustüüpide mitmekesisus eesti keeles

tuba-toa (b:0), tõmbama-tõmmata (b:m), varvas-varba, kaebama-kaevata (b:v)
jalg-jala, kurg-kure, nuga-noa (g:0), selg-selja, järg-järje (g:j)
rida-rea, luud-luua (d:0), sada-saja (d:j), känd-kännu (d:n), kallas-kalda (d:l), kord-korra (d:r)
käsi-käe (s:0), vars-varre (s:r)
kott-koti, laht-lahe, ütlema-öelda (t:0), pilt-pildi (t:d)
kukkuma-kukun, käskima-käsin, õhk-õhu (k:0), telk-telgi, pank-panga, virk-virga (k:g)
lipp-lipu (p:0), karp-karbi, tulp-tulbi (p:b)

Analüüsides eelnevas näites esinevaid vastavusi tugeva ja nõrga astme häälikute vahel, selgub, et vastavus tugeva ja nõrga astme häälikute vahel ei ole üksühene. Üks võimalus mitmesustest vabanemiseks oleks iga astmevahelduse tüübi erinev tähistus süvaesituses. See aga ähmastab sõnatüve sõnastikuesitust. Mõistlikum lahendus oleks erinevate astmevaheldustüüpide kontekstide eristamine, kuid kas see on kõikide astmevaheldustüüpide puhul võimalik? Töös Uibo 1999 on üksikasjalikult analüüsitud kõiki tugeva astme häälikuid: nende kõikvõimalikke vastavusi ja kontekste, milles sellised vastavused esinevad. Reeglid kujunesid välja praktilisel reeglite testimisel. Kõikidel juhtudel osutus kontekstide eristamine küll

võimalikuks, kuid mõnel juhul läksid kontekstid tõepoolest väga detailseks.

Vaatleme näiteks vastavusi *G:0* ja *G:j*. Kui G-le eelneb vahetult vokaal, siis ta nõrgas astmes kaob:

G:0 ⇐ Vok _ %$:;

Kui aga **g**-le eelneb **l** või **r**, siis ühetähelisest vasak- ja parempoolsest kontekstist ei piisa:

LC = l, RC = a	LC = r, RC = e
*jalg-jal***a** (g:0)	*kurg-kur***e** (g:0)
*selg-sel***ja** (g:j)	*särg-sär***je** (g:j)

Uurime, kas *l*-le või *r*-le eelneva vokaali kaasamine konteksti võimaldab neid astmevahelduse tüüpe eristada. Kirjutame tabeli 1 vasakpoolsesse tulpa *g:0*-tüüpi sõnad ning parempoolsesse tulpa g:j-tüüpi sõnad ning toome välja kahetähelise vasakpoolse ning ühetähelise parempoolse konteksti. Grupeerime ühesuguse esisilbi vokaaliga sõnad.

Tabel 1. Astmevaheldustüüpide g:0 ja g:j eristamine vasak- ja parempoolse konteksti põhjal, kui vahetu vasakpoolne kontekst on "l" või "r"

Esisilbi vokaal	g:0	LC	RC	g:j	LC	RC
A	jalg-jala	a l	a			
	palge-pale	a l	e			
	halg-halu	a l	u			
E	veerg-veeru	e r	u	selg-selja	e l	a
				telg-telje	e l	e
I	hiilgama-hiilata	I l	a			
	kiirgama-kiirata	I r	a			
	kirg-kire	I r	e			
	viirg-viiru	I r	u			
U	sulg-sule	u l	e			
	kurg-kure	u r	e			
	sulgida-sulib	u l	i			
	sulg-sulu	u l	u			
	urg-uru	u r	u			
Õ	võlg-võla	õ l	a			
	sõrg-sõra	õ r	a			
	sõlg-sõle	õ l	e			
Ä	sälg-sälu	ä l	u	nälg-nälja	ä l	a
				märg-märja	ä r	a
				jälg-jälje	ä l	e
				järg-järje	ä r	e
Ü	nülgida-nülib	ü l	i	hülgama-hüljata	ü l	a
	pürgida-pürib	ü r	i	külg-külje	ü l	e

Tabelist 1 selgub, et esisilbi vokaali *a, i, u* ja *õ* puhul ei pea paremat konteksti ehk tüvevokaali arvestama: astmevahelduse tüübiks on siin g:0. Kastiga ümbritsetud ja seega problemaatilised on sõnad, millel esisilbi vokaaliks on *e, ä* või *ü*. Siin aga otsustab tüvevokaal, kumma astmevaheldustüübiga on tegemist. Seejuures saab kõik kolm juhtu ühte reeglisse (Uibo 1999) kokku võtta: "Kui esisilbi vokaal on *e, ä* või *ü* ja tüvevokaal on *i* või *u*, siis tugeva astme g nõrgas astmes kaob, kui tüvevokaal on *a* või *e*, siis vastab g nõrgas astmes j-le"

G laadivaheldust käsitlevad kahetasemelised reeglid on siis järgmised:

"AV 11: g kadu" !jalg-jala, kirg-kire
G:0 ⇔ Vok _ (Vok|h) %$: ;
[a | i | õ | u] [l | r] _ (Vok) %$: ;
[e | ä | ü] [l | r] _ [i | u] %$: ;
"AV 17: g:j" !märg-märja, hüljes-hülge
G:j ⇔ [e | ä | ü] [l | r] _ [a | e] (S:) %$: ;

Iseäranis palju on neid laadivahelduse tüüpe, milles tugevat astet esindab d. Järgnevast selgub, miks joonisel 3 esitatud reegel oli niivõrd keeruline. Erinevaid nõrga astme variante on siin viis: *0, j, l, n* ja *r*. Kolmel viimasel juhul on tegemist assimilatsiooniga ning need astmevahelduse tüübid on väga selgelt vasaku kontekstiga määratud: *d* assimileerub selle konsonandiga hulgast {l, n, r}, mis talle vahetult eelneb. Sarnaselt *g:0* ja *g:j*-ga on problemaatiline *d:0* ja *d:j* kontekstide eristamine. Kui *d*-le eelneb **pikk silp**, siis *d* nõrgas astmes kaob (*laud-laua, siirdama-siirata*), kuid **lühikese silbi** järel võib *d* kas kaduda või muutuda *j*-ks (*vedada-vean, madu-mao, rida-rea, õdeõe, küdeda-köeb* (*d:0*), aga *rada-raja, sadu-saju, koda-koja, sõdasõja* (*d:j*)). Analoogilisel kontekstide võrdlemisel, nagu tegime tabelis 1, ilmneb, et (Uibo 1999):

- esisilbi vokaalide[2] *e, i, u, ü* puhul on vastavus alati *d:0*;
- esisilbi vokaali *o* puhul on vastavus alati *d:j*;
- esisilbi *õ* ja tüvevokaali *e* puhul esineb kadu (*õde-õe*), aga tüvevokaalide *a* ja *u* puhul muutub *d* nõrgas astmes *j*-ks (*sõda-sõjasõjust*);
- esisilbi *a* ja tüvevokaali *a* puhul alati vastavus *d:j*, kuid tüvevokaal *u* sunnib lisaks vaatama sõnaalguse konsonanti – võiks öelda, et

[2] Esisilbi vokaali võtame algvormist, mitte arvestades vokaali madaldumist nõrgas astmes (nt *rida-rea* – esisilbis *i*, *kududa-koon* – esisilbis *u*).

siin on tegemist väga ebareeglipärase juhtumiga, kuna fonoloogiliselt kujult väga sarnased sõnad muutuvad erinevalt, vrd *sadu-saju, rada-(radadelt e) rajult* ning *madu-mao, ladu-lao, kadu-kao*. Siin tuleb sisuliselt kõik üksikjuhtumid panna reeglisse, mis pole üldiselt soovitav. Kuid uusi sõnu *d:j*-tüüpi tõenäoliselt ei tule, kuna laadivaheldus on tänapäeva eesti keeles muutunud ebaproduktiivseks, st laadivaheldus taandub vananevast mitteaktiivsest sõnavarast ning uued keelde tulevad sõnad ei ühine laadivahelduslike tüüpidega (Hint 1997).

Eesti keeles esineb ka prosoodiline astmevaheldus, mis kirjapildis ei avaldu, kuna teise ja kolmanda välte erinevus ilmneb kirjapildis ainult klusiilidel. Kuivõrd see astmevahelduse tüüp kirjapilti ei mõjuta, jäi astmevaheldusliikkus nende tüvedes markeerimata. Kui sõnastike süsteem aga Ü. Viksi klassifikatsioonile (Viks 1994a) kohandada, tuleks välde siiski markeerida ja tüübistik ümber vaadata.

3.2.2. Fonotaktika reeglid

Fonotaktika reeglid ütlevad, millised häälikujärjendid sobivad eesti keele sõnadeks ja millised mitte (tähendusega pole siin midagi tegemist). Nad keelavad ära teatud häälikujärjendid, mõne hääliku esinemise järgsilbis jne. Fonotaktika reegleid on üldiselt rohkem, kui käesolevas reeglistikus formaliseeritud. Kahetasemelised reeglid pööravad tähelepanu ainult nendele häälikujärjenditele, mis võivad morfoloogiliste formatiivide liitumisel või tüvevahelduse (eeskätt astmevahelduse) mõjul tekkida.

Allikast (Hint 1978) leidsin hulga seaduspärasusi, mis on kahetasemeliste reeglitena vormistatud.

1. Häälik *h* saab olla pearõhulise silbi lühikese vokaali järel ja sõna alguses ning ainult seal. *h* esinemise piirangutest on tekkinud vokaali lühenemine vormides *maha, töhe, öhe, sohu, pähe.*

Samal ajal toimib ka teine fonotaktika reegel:

2. Järgsilpides pole *o, ö, ä* lubatud ning toimub teisenemine *o→u, ö→e, ä→e*.
3. Võimalikud on kõik pikad vokaalid, piiranguid on diftongidele:
   i võib diftongi 2. komponendiks olla suvalise 1. komponendi korral,
   u ei saa olla *ö, ü* järel,

e saab olla *a, o, õ, ä* järel,
a võib olla ainult *e* järel.

Sellest tulenevad häälikumuutused astmevaheldusliikes sõnades:

tugi-tue →toe* ***ue→oe***
tuba-tua→toa* ***ua→oa***
rida-ria→rea* ***ia→ea***
süsi-süe→söe* ***üe→öe***

Saame reegli

```
"Vokaali madaldumine"
KorgeVok:MadalVok ⇔ Algus _ LV: [a|e|i:|u:](l) %$: ;
            Algus Vok LV: _ %$: ;
                  where KorgeVok in (u ü i)
                        MadalVok in (o ö e)
                  matched ;
```

4. *lumi-lumd*→lund*
   leem-leemt→leent* ***m→n*** enne *t*-d või *d*-d
5. *lodi-lotji*→lotje*
   ori-orji→orje*
   osi-osji→osje*
   Siin keelab fonotaktika häälikujärjendi *ji*, mis morfoloogilise reeglipära järgi peaks tekkima.
6. *laps-lapst*→last*
7. *uks-ukst*→ust* – kirjakeel on vastu tulnud häälduspärasusele.
8. Vokaal *a, e* või *u* kaob *l, m, n, r* eest, kui lisandub tüvevokaal:
   sõber-sõbra *kannel-kandle*
   vaagen-vaagna *koorem-koorma*
   Vahevokaal tuleb sõnastikuesituses märkida suurtäheliselt, kuna kaduva vahevokaali markeerimata jätmisel oleks kaoreegel hakanud toimima ka sellistele laensõnadele nagu *kanal-kanali (kanli*), pinal-pinali (pinli*)* jt.
9. "*oo+i, öö+i →õi*" *!sööma-sõi, jooma-jõi*

3.2.3. Morfeemipiiridel toimivad reeglid

Ü. Viksi "Väikese vormisõnastiku" grammatikast (Viks 1992) võib leida kaks üldisemat laadi morfofonoloogilise distributsiooni reeglit.

1. Pikk vokaal lüheneb enne *i*-ga algavat formatiivi (***kuu-kuid***).
2. *i→e* enne *i*-ga algavat formatiivi (***naaskli-naaskleid***, *kauni-kauneim*).

3. Reegel "Pikk madal vokaal kõrgeneb *a* ja *e* ees" (*oo+a->uua, öö+a->üüa*) on näitena toodud Ü. Viksi dissertatsioonis (Viks 1994a).
4. Samast (Viks 1994a) on leitud ka konsonantidevahelise siirdevokaali reegel.

Praktilise testimise käigus on osutunud kasulikeks järgmised morfeemipiiridel kehtivad reeglid.

1. "Morfeemipiiril kaks *s*-i saavad üheks" (*inimes+sse=inimesse*).
2. "Morfeemipiiril kaks *d*-d saavad üheks ning *dt* saab *t*-ks" (*and+da=anda, and+tud=antud*).
3. "Käskiva kõneviisi formatiivi *-gu* või *-ge* alguse *g* läheb *k*-ks helitu konsonandi järel või NA-tüves *a* järel" (*tehke, andke, võtke; hakake*).
4. "Formatiivi alguse *d* läheb *t*-ks helitu konsonandi järel" (*seis+da=seista*).
5. "Tüvevokaal kaob *us*-liite ees" (*raske+us=raskus, puhta+us=puhtus*).
6. "*e→i* enne tegijanime tunnust '*ja*'" (*tegija, nägija*).

Kõik nimetatud reeglid on realiseeritud kahetasemeliste reeglitena.

3.2.4. Ortograafiareeglid

Süsteemis on realiseeritud järgmised õigekirjareeglid.

1. Kaashäälikuühendis kirjutatakse kõik häälikud ühekordselt (Erelt 1997: 9), nt *kukkru*→kukru, kristallne*→kristalne*.
2. Kolm või enam ühesugust häälikut kõrvuti eraldatakse loetavuse huvides sidekriipsuga (Erelt 1997: 79). (*plekk-katus, jää-äär*).
3. Silbi (siis ka sõna) lõpus kirjutatakse *j* asemel *i* (Erelt 1997: 7) (*kirj*→kiri, purj*→puri*).

3.2.5. Reeglite kirjutamisel tekkinud probleemid

Reeglite koostamisel tuli kokku puutuda mitmete probleemidega, mis osaliselt johtuvad kasutatava formalismi iseärasustest, osaliselt eesti keele morfoloogilise süsteemi keerukusest (Uibo 1999).

1. Tähistusprobleemid. Süvaesitusse uute sümbolite lisamisel oleks mõistlik silmas pidada, et sümbol oleks tähenduselt nähtusega kuidagi samastatav.

- Astmevahelduse korral oleks soovitav süvaesituse sümboli ühtivus tugeva astme allomorfiga. Üksikasjaliku kontekstide analüüsi teel õnnestus seda põhimõtet ka järgida.
- Vokaali kadu ilmneb mitmel juhul, kusjuures kaduvate vokaalide hulgad on teatava ühisosaga, kuid täielikult ei kattu:
  a) Vahevokaali kao reeglisse on haaratud *a, e, u*. Kaduv vokaal tuli siin kindlasti markeerida (vastavate suurtähtedega A, E, U), kuna sarnane häälikuline kuju seda nähtust alati ei põhjusta.
  b) Pikad vokaalid *aa, ee, oo, uu, ää, öö, üü* lühenevad enne *i*-ga algavat formatiivi. Pika vokaali traditsiooniline tähistus, näiteks häälduse märkimisel, on koolon. Kahetasemelistes reeglites on koolon aga süva- ja pindsümboli eraldajaks reegli vasakul poolel. Selline tähistus oleks küll lubatav, kui kirjutada kooloni ette protsendimärk, aga reegli loetavuse seisukohalt on see ikkagi halb. Mingi teine täht tekitaks ka segadust. Seetõttu sai valitud pika vokaali teise komponendi süvasümboliks punkt. Vaikimisi pikendab punkt eelseisvat vokaali, mitmuse formatiiv *i* tekitab aga pindesitusse tühisümboli.
  c) Tüvevokaal (*a, e, i, u*) kaob tüvemitmuse vokaali eest. Kuna siin on parem kontekst väga selgelt piiritletud, siis võib need vokaalid jätta sõnastikus väiketähelisteks.

Kui süvaesituses markeeritavaid nähtusi on palju, siis pole tähistuse mõtestatuse idee alati teostatav, veelgi halvem – klaviatuurisümbolid võivad otsa lõppeda. Kahetasemelised reeglid võtavad aga süvaesitusest ühe sümboli korraga, seega pikemaid tähiseid ei saa kasutada. Teine probleem on see, et regulaarsete avaldiste süntaksis on suurem osa mittetähelistest sümbolitest juba kasutusel. Nende tähenduse saab protsendimärgiga küll tühistada, aga sellega halveneb reeglite loetavus.

2. Eesti keeles on palju astmevahelduse tüüpe, kusjuures vastavus tugeva ja nõrga astme allomorfide vahel pole üksühene. Mõnel juhul on kontekstid, milles esineb üks ja teine vastavus, raskesti eristatavad (vrd *jälg-jälje* ja *jalg-jala, ladu-lao* ja *sadu-saju*).

3. Häälikumuutuste käsitlemine reeglitega on mõnikord raske. Nt *lööma-lõi, jooma-jõi*, aga *töö-töid* (mitte *tõid**), *soo-soid* (mitte *sõid**). Samal ajal pole meil reeglis võimalik kasutada infot sõnaliigi

kohta. Et seda nähtust reegliga käsitleda, peab parema konteksti sõnastikuesitus olema erinev: *tö.+i+d* → *töid*, aga *to.+id* → *tõid*. See tähendab, et verbi formatiiv jääb algosadeks lahutamata.

4. Puudub võimalus ühe reegliga käsitleda tervet segmenti. Näiteks teisendused *jooma-juua* (*oo→uu*), *saba-sappa* (*b→pp*) tuli ära teha kahe reegliga. Niisugusel juhul peab aga vältima olukorda, kus reeglid kasutavad vastastikku teineteise vasaku poole pindsümbolit kontekstis – see tekitab lõpmatu rekursiivse pöördumise. Kindlaim viis seda vältida on süvaesituse kontekstide kasutamine, kui vähegi võimalik.

5. Kui mitmed reeglid kasutavad sulundioperatsiooni (*), siis võtab automaatide korrutise leidmine aega ning ühendautomaat läheb väga suureks. Nt astmevahelduse käsitlemisel üldiste reeglitega nagu

D:0 ⇔ _ (Alfa) * $:;

kulus XRCE programmil *lexc* ühendautomaadi koostamiseks 10 minutit (Uibo 1999). Sisuliselt sama, kuid täpsustatud, sulundioperatsiooni vältivate kontekstidega reeglitekogu ühendamiseks sõnastikuga kulus vaevalt 1 sekund. Konteksti täpsustamine osutub tüübiloendite võrdlemisel tihti võimalikuks, kuigi esialgu on mugavam kirjutada üldisema kontekstiga reegleid. Koskenniemi (Koskenniemi 1997) märgib sama probleemi lõplikel automaatidel põhineva süntaksi puhul.

3.3. Eesti keele kahetasemelise morfoloogia sõnastike süsteem

Sõnastike süsteemi ülesandeks kahetasemelises mudelis on defineerida kõik morfotaktiliselt võimalikud morfeemijadad. Hetkel ulatub juursõnastike kirjete arv vaid 350ni, kuid juba nii väikestki hulka erinevaid leksikaalseid tähendusi kasutades on sõnastike süsteem võimeline genereerima üle 20 000 lihtsõnavormi ning lõpmatu arvu liitsõnu (kuna liitsõnakomponentide arv pole piiratud).

Kirjete grupeerimine erinevatesse sõnastikesse on määratud nende võimaliku kombinatoorikaga, so. millele antud sõnaosa eelneb ja millele järgneb. Iga jätkamisviit osutab ühele või mitmele sõnastikule, millest otsida sobivat sõnajuppi jätkamiseks. Selline viitstruktuur väldib kordusi: näiteks pole meil tarvis kahte eraldi sõnastikku nimi- ja omadussõnade käändelõppude jaoks.

Eesti keel on rikas tüvevahelduste poolest. Tüvevahelduste kirjeldamisest reeglitega oli juttu jaotises 3.2. Teine võimalus on käsitleda tüvevaheldusi sõnastikega. Kuigi mõned tüvevahelduse tüübid on tänapäeva eesti keeles muutunud ebaproduktiivseiks, loetakse isegi paari-kolme tüve poolt viidatavat jätkusõnastikku ökonoomsemaks kui mitme tüvevariandi kirjutamist sõnastikku (Koskenniemi 1983).

Keeleajalooliselt on eesti keele tüvevaheldusmallid olnud suurel määral morfofonoloogiliselt tingitud (Viks 1994a). Astmevaheldust mõjutas sõnavormi silbistruktuur: kahesilbilistes tüvedes põhjustas kinnine teine silp nõrga astme tekke, lahtise silbi korral kujunes tugev aste. Lõpuvaheldus oli seotud kindlate vormidega: kahetüvelistel sõnadel oli teatud vormides kasutusel vokaaltüvi, teatud vormides konsonanttüvi. Tänaseks on keeles toimunud palju häälikumuutusi, mille tagajärjel tüvevaheldusi põhjustanud tegurid võivad olla kadunud, kuid vaheldusmallid ise on säilinud. Seetõttu on tüvevaheldusmallid, nagu ka astmevaheldusmallid (jaotis 3.2.1.) saanud suurel määral sõna individuaalseks iseärasuseks ning neid ei saa enam häälikulisel ümbrusel põhinevate reeglitega kirjeldada.

Lemmatüvede levik paradigmas on verbil ja noomenil erinev (Viks 1990): verbide lemmatüvede levik langeb alati kokku vastavalt kas tugeva või nõrga astme levikuga. Noomenitel tuleb lemmatüvi tavaliselt esile ainult ainsuse nimetavas, kõigis teistes vormides figureerib muutetüvi, sõltumata sellest, kas ta on seal tugevas või nõrgas astmes või astmevaheldusetа. Kuid umbes 300 noomeni puhul on levik teistsugune: nendes astmevaheldusliкes muuttüüpides, kus ainsuse nimetav on nõrgas astmes, kandub lemmatüvi üle ka teistesse nõrgaastmelistesse vormidesse. Kuid *sõber-* ja *padi-* tüübis seevastu on ainult ainsuse nimetav nõrgas astmes ning lemmatüvi teistesse nõrgaastmelistesse vormidesse ei kandu.

Lemma- ja muutetüvede paiknemine erinevate sõnade puhul
hammas *– hamba –* ***hammast*** *–* ***hammaste*** *– hambaid*
liige *– liikme –* ***liiget*** *– liikmete – liikmeid*
aga: ***sõber*** *–* <u>*sõbra*</u> *– sõpra – sõprade – sõpru*

Formatiivide variandid on algselt samuti olnud morfofonoloogiliselt tingitud, sõltudes formatiivi asendist sõnavormi silbistruktuuris, silbi rõhulisusest ja eelnevatest häälikutest. Keeleajaloolised arengud on sellegi sõltuvuse paljudel juhtudel kaotanud ja jätnud formatiivivariantide valiku sõna individuaalseks iseärasuseks.

3.3.1. Käändsõnade sõnastikud

Morfotaktika üldreegel (Viks 1994a) noomenivormide moodustamiseks on järgmine:

Tüvi + arv + kääne

Noomenite puhul on olemas teatavad üldised reeglid selle kohta, millistest tüvedest missugused käänded moodustatakse (tabel 2).

Tabel 2. Noomeni paradigmas tavaliselt üksteisest moodustatavad käänded (Viitso 1990)

Ainsuse omastav +	*-sse, -s, -st, -le, -l, -lt, -ks, -ni, -na, -ta, -ga, -d*	ainsuse käänded sisseütlevast kaasaütlevani ning mitmuse nimetav
	-te või *-de*	mitmuse omastav
	(-te või *-de)* + *-sse, -s, -st, -le, -l, -lt, -ks, -ni, -na, -ta, -ga*	mitmuse käänded sisseütlevast kaasaütlevani

Selles rühmas on vormimoodustus aglutinatiivne: ainsuse omastavast saadakse kõik ülejäänud vormid (mitmuse tunnuse ja) käändelõpu lisamisel. Nimetatud reeglipärast lähtudes on kirjeldatud sõnastikud S1 (käändelõpud sisseütlevast kaasaütlevani) ning S2 (*de*-mitmuse käänded). Ainsuse käänded sisseütlevast kaasaütlevani saadakse kõikide muuttüüpide puhul kirjest

+Sg:<ainsuse omastava tüvi> S1;

ning *de*-mitmuse käänded (tüüpidel, mille mitmuse tunnus on *-de*, mitte *-te* ja mitmuse osastava lõpp *-sid*) kirjest

+Pl:<ainsuse omastava tüvi> S2;

kusjuures sõnastikud S1 ja S2 on järgmised:

LEXICON S1
! sisseütlevast kuni kaasaütlevani
+Ill:+sse #;
+In:+s #;
+El:+st #;
+All:+le #;
+Ad:+l #;
+Abl:+lt #;
+Ter:+ks #;
+Trl:+ni #;
+Es:+na #;
+Ab:+ta #;
+Kom:+ga #;

LEXICON S2
! de-mitmuse käänded
+N:+d #;
+G:+de #;
:+de S1;
+P:+sid #;

Ühe põhivormina toimib ka mitmuse osastava lühike vorm (*sõna-sõnu-sõnus, sõnust* jne.), kuid nõrgeneva astmevaheldusega tüüpide puhul selline vormimoodustusmuster ei toimi. Seal tuleb mitmuse osastava tugev aste asendada muudes tüvemitmuse vormides nõrga astmega (*jalg-jalgu-jalus, jalust* jne.).

Tüvemitmuse lõppude sõnastikud on S4 ja S5, kusjuures S4 = S5 ∪ {osastava, rajava, oleva käände lõpud} ja S5 = {käändelõpud sisseütlevast saavani}, kuna üldiselt pole viimases neljas käändes tüvemitmus kasutatav (Kaalep 1999: 27). H.-J. Kaalepi praktika eestikeelsete tekstide analüüsimisel näitab veel, et *lugemik*-tüübis on tüvemitmus kasutatav kõigis käänetes, *i*-mitmusega sõnadel on vokaalmitmus võimatu kahe viimase käände puhul ning seitsme erandsõna puhul on vokaalmitmus võimatu kolmes viimases käändes. Vastavalt on organiseeritud ka kahetasemelise morfoloogia sõnastikud.

Nimisõnatüvikute sõnastikus on hetkel sadakond kirjet. Jätkusõnastike nimeks on mõnedel juhtudel võetud tüüpsõna nimi (KUU, AASTA, JALG jt.), tüvelõpuvaheldusega sõnade jätkusõnastiku nimes kajastub vahelduse tüüp (A, U, 0-DA, S-KSE) ning astmevaheldusliikel tüüpidel on sõnastiku nimesse märgitud, kas tegemist on tugevneva või nõrgeneva astmevaheldusega (LV/N – nõrgenev laadivaheldus, T – tüvelõpuvahelduseta tugevnev astmevaheldus).

Astmevaheldus kombineerub teiste tüvemuutustega (Hint 1997: 49):

1) vokaal- ja konsonanttüve vaheldumine (nt I/N – muutetüves tüvevokaal *i*, nõrgenev astmevaheldus, E/T – muutetüves tüvevokaal *e*, tugevnev astmevaheldus);
2) tüvevokaali teisenemine (nt I-E-T/N – lemmatüve vokaal *i*, muutetüvel *e*, osastava lõpp *t*, nõrgenev astmevaheldus);
3) *me*-sufiksi esinemine–puudumine (ME/T).

Tüvelõpumuutused ei tarvitse noomenite puhul "käia ühte jalga" astmevaheldusega (Viks 1990), seetõttu võib astmevahelduslike noomenitüüpide jätkuleksikonide arv ulatuda kuni järgmise korrutiseni:

astmevaheldusliikide arv (nõrgenev, tugevnev) x tüvelõpumuutuste arv

Omadussõna paradigma erineb nimisõna omast võrdlusastmete olemasolu poolest. Igast omadussõnast saab moodustada kesk- ja ülivõrde, kusjuures need käänduvad nagu tavalised *tänav*-tüüpi noomenid (muutetüve vokaal *a*, vähemalt 3 silpi). Keskvõrre (ja

ühtlasi ka *kõige*-ülivõrre) moodustatakse ainsuse omastava tüvest, i-ülivõrre ehk lühike ülivõrre tüvemitmuse tüvest (kui see esineb). Kui tüvi lõpeb *i*-ga, toimib siin reegel "*i→e* enne *i*-ga algavat formatiivi" (*kallis-kallim-kalleim*). Kui tekib häälikujärjend "*ji*" (*tühi-tühjem-tühjim**), siis lühikest ülivõrret moodustada ei saa, kuna järjendit "*ji*" eesti keele fonotaktika ei luba. Samuti ei ole lühikest ülivõrret *i*-tüvelistel omadussõnadel (*karm-karmim-kõige karmim*).

Omadussõnatüübi "õnnelik" jätkuleksikon IK/O

```
LEXICON IK/O          !õnnelik
+A:0 IK';
+A+Cmp:u$+m A';
+A+Spr:$+em A';
kus:+us E-SI;
kult+Adv:u$+lt #;
```

Algvõrde kõik käänded moodustatakse nagu *lugemik*-tüübis (viidates sõnastikule IK'). Kesk- ja ülivõrre moodustatakse nõrga astme tüvest ja need käänduvad nagu *tänav* (minisõnastik A'). Igast seda tüüpi sõnast saab *us*-liite abil tuletada nimisõna (käändub nagu *küsimus*) ning *lt*-liite abil määrsõna.

Asesõnade hulgas on väga palju erandlikult käänduvaid sõnu. Seepärast on nende jaoks loodud sõnastikud MA, TA, ME, NAD, SEE, NEED, ISE, KES jt. Need asesõnad, mis pole erandlikud, käänduvad nimisõnatüüpide eeskujul. Arvsõnad *üks, kaks* ja *kümme* viitavad erandtüüpide sõnastikele YKS ja KYMME, teised arvsõnad kasutavad nimisõnade jätkusõnastikke.

3.3.2. Pöördsõnade sõnastikud

Verbi morfotaktiline struktuur on üsna keeruline ja hõlmab järgmisi grammatilisi kategooriaid (Viks 1980): kõneviis, aeg, tegumood, arv, isik, kõneliik. Seejuures eitav kõneliik ning täis- ja enneminevik moodustatakse abisõnadega (vastavalt *ei, ära* jne ning abiverb *olema*).

Kõikide grammatiliste kategooriate tunnused ei pea olema fonoloogilisel tasandil esindatud. Mõned neist on koguni teineteist välistavad, näiteks arv ja isik saavad esineda ainult isikulises tegumoes. Iga grammatilise kategooria kõige tavalisemad variandid – isikuline tegumood, kindel kõneviis, olevik ja jaatav kõneliik – on tunnuseta.

Sõnavormi *naernuksime* liigendus morfeemideks

0	naer	+nu	+0	+ksi	+me
↑	↑	↑	↑	↑	↑
kõneliik tüvi	aeg	tegumood		kõneviis	arv ja isik
(jaatav)	(minevik)	(isikuline)		(tingiv)	(mitmuse 1. isik)

Ü. Viksi klassifikatsiooni kohaselt võib verbid üldisemalt liigitada kolme pöördkonda (Viks 1980). Verbil võib olla ülimalt neli erinevat tüve:

1) *ma*-tegevusnime tüvi;
2) *da*-tegevusnime tüvi;
3) isikulise tegumoe kindla kõneviisi oleviku tüvi;
4) umbisikulise tegumoe tüvi.

Ü. Viks annab kõigi kolme pöördkonna jaoks **analoogiareeglid** (Viks 1980), mille alusel saab kindlaks teha, milline tüvi esineb millistes muutevormides.

Nt I pöördkonnas (tüübid *elama* (üle 5000 sõna), *õppima* (üle 2000 sõna), *naerma* (44 sõna), *hakkama* (üle 700 sõna)) on analoogiarühmad järgmised:

1) *-ma, -mas, -mast, -maks, -mata, -vat, -v, -sin, -sid, -s, -sime, -site*;
2) *-da, -des, -ge, -gu, -gem, -nud, -nuksin, -nuksid, -nuks, -nuksime, -nuksite, -nuvat*;
3) *-n, -d, -b, -me, -te, -vad, -0* (imp pr sg 2, ind pr ps neg), *-ksin, -ksid, -ks, -ksime, -ksite*;
4) *-tud, -ti, -taks, -tuks, -tavat, -tuvat, -tagu, -tama, -ta, -tav.*

Kõige produktiivsem verbitüüp *elama* (täiesti muutumatu tüvega) on kirjeldatud jätkusõnastikus V1, mis omakorda viitab sõnastikele V10–V13. *ele*-lõpuliste tegusõnade muutumist kirjeldavad sõnastikud V10–V14 (Sõnastikule V14 viidates saab moodustada umbisikulise tegumoe tüvest lähtuvad vormid, kui umbisikulise tegumoe tunnus on *-t*).

I pöördkonda kuuluvad ka tugevneva ja nõrgeneva astmevaheldusega tüübid. Need kasutavad samuti jätkusõnastikke V10–V14. Erinevalt noomenitest on verbidel lemmatüvi (*ma*-tegevusnime tüvi) alati tugevas astmes.

Tugevnevat astmevaheldust käsitlevad jätkusõnastikud TUG (tüvelõpuvahelduseta, nt *hakkama*), TUG-PROS (tüvelõpuvahelduseta prosoodiline astmevaheldus, nt *pöörama*), LE-EL/T (tüvelõpuvaheldus *le-el*, nt *vaidlema*).

Nõrgeneva astmevaheldusega tüübid kasutavad jätkusõnastikke N (tüvelõpuvahelduseta, nt *õppima*), N-PROS (tüvelõpuvahelduseta prosoodiline astmevaheldus, nt *istuma*), 0-A (tüvelõpuvaheldus *0-a*, nt *laulma*), 0-E (tüvelõpuvaheldus *0-e*, nt *kuulma*), TAHTMA, TEGEMA, 0-A/N (tüvelõpuvaheldus *0-a, tt-t, pp-p, t-d* või prosoodiline, *da*-infinitiiv *-a*, nt *saatma*), JÄTMA, PIDA-SIN, PID-IN, I-0/N (umbisikulises tegumoes konsonanttüvi, nt *käskima*).

Verbidel on kolme tüüpi lihtmineviku lõppe, mis on vastavalt kirja pandud sõnastikes Ipt1 (*-sin, -sid, -is, -sime, -site, -sid*), Ipt2 (*-in, -id, -i, -ime, -ite, -id*) ja Ipt3 (*-sin, -sid, -s, -sime, -site, -sid*). Käskiva kõneviisi muutelõpud (*-gu, -gem, -ge*) on sõnastikus Imp, tingiva kõneviisi oleviku lõpud sõnastikus Knd-Pr ning tingiva ja kaudse kõneviisi mineviku lõpud sõnastikus Knd, Kvt-Pt.

3.3.3. Muutumatute sõnade sõnastikud

Kuivõrd side- ja hüüdsõnade kirjeldus on triviaalne – kõik juursõnastiku kirjed on kujul sõna+liik:sõna # –, on nad esialgu sõnastike süsteemi lisamata jäetud.

Ka kaas- ja määrsõnad on reeglina käändumatud, kuid mõned kohamäärsõnad käänduvad sise- või väliskohakäänetes. Selline osaline käändumine on jätkusõnastikuga kirjeldatud:

```
Väljavõte kaassõnade juurtesõnastikust
ning väliskohakäänete sõnastik LE-LT
LEXICON Kaassõna
.....
järel LE-LT;              !LE-LT genereerib kaassõnad järele, järel, järelt
järgi+K:järgi #;
kaudu+K:kaudu #;
korral+K:korral #;
kõrval LE-LT;
mööda+K:mööda #;
peal LE-LT;
pool LE-LT;
.....
LEXICON LE-LT             !väliskohakäänetes käänduvate kaas- ja määr-
                          sõnade jaoks
e+K:e #;
+K:0 #;
t+K:t #;
```

Mõned määrsõnad võivad olla ka liitsõna eeskomponendiks. Määrsõnade liitumist käsitleb jätkusõnastik PREFIX:

```
LEXICON PREFIX
+Adv: #;
+:& Verbituletis;
```

3.3.4. Produktiivne sõnatuletus

Tuletusliigenduse puhul kirjeldavad morfotaktika reeglid ka tüve tuletusstruktuuri, so. millised derivatsioonisufiksid ja millises järjestuses võivad juurele liituda. Igast verbist moodustuvad sellised tuletised on näha tabelis 3.

Tabel 3. Verbituletised

Tuletussufiks	Näide (*lugema*)	Tuletise sõnaliik
-ja	lugeja	nimisõna
-mine	lugemine	nimisõna
-v	lugev	omadussõna
-tav	loetav	omadussõna
-nud	lugenud	omadussõna
-tud	loetud	omadussõna
-nu	lugenu	nimisõna
-tu	loetu	nimisõna

Omadussõnast saab tuletusliite *-us* abil moodustada nimisõna ning liidete *-sti* ja *-lt* abil moodustada määrsõnu. Kahe viimase kasutamine on tegelikult sõnati erinev: mõne omadussõna puhul on võimalikud nii *sti-* kui *lt*-tuletis, mõne sõna puhul ainult üks või teine.

Määrsõnade tuletatavuse poolest erinevad grupid ei pruugi üldiselt ühtida morfoloogiliste muuttüüpidega (vrd tabel 4 *halb*, *nõrk*, *pikk* – sama muuttüüp, kuid erinevad tuletised, samuti *kauge* ja *kõrge, kuri* ja *tühi*). Ka Ü. Viksi seisukoht on, et enamasti peab konkreetse sõna juures ära märkima, kas talle mingi tuletusliide sobib või ei (Viks 1994a).

Tabel 4. Omadussõnadest tuletatavad määrsõnad

Omadussõna	-sti	-lt	Muuttüüpi määravad tunnused
halb	halvasti	–	nõrgenev AV, lõpuvaheldus 0-a
kauge	–	kaugelt	tüvemuutusteta
kuri	kurjasti	kurjalt	nõrgenev AV, lõpuvaheldus i-ja

Omadussõna	-sti	-lt	Muuttüüpi määravad tunnused
kõrge	kõrgesti	kõrgelt	tüvemuutusteta
noor	–	noorelt	nõrgenev AV, lõpuvaheldus 0-e
nõrk	nõrgasti	nõrgalt	nõrgenev AV, lõpuvaheldus 0-a
pikk	–	pikalt	nõrgenev AV, lõpuvaheldus 0-a
suur	suuresti	suurelt	nõrgenev AV, lõpuvaheldus 0-e
tühi	–	tühjalt	nõrgenev AV, lõpuvaheldus i-ja

Probleem on ka astmevahelduslike verbide tuletistega. Nimelt on verbi algvorm alati tugevaastmeline, kuid verbituletise morfoloogiline info peab sisaldama tuletise, mitte alusverbi tüve. Sõnastikteisendaja korjab verbitüve juurest alguses üles tugevaastmelise tüve ja sõnaliigi V (verb), kuid tuletusliite leidmisel võib osutuda, et tegemist on hoopis nimi- või omadussõnaga, mille juur on nõrgaastmeline. Nii nõrgeneva kui tugevneva astmevaheldusega sõnadel liitub osa tuletussufiksitest nõrgaastmelisele tüvele. Seetõttu peab astmevahelduslike verbide tüved kaks korda – nii tugeva- kui nõrgaastmelisena sõnastikku panema.

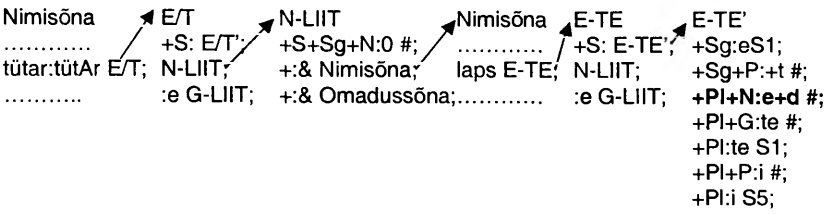
3.3.5. Liitsõnade moodustamine

Vastavalt eesti keele õigekirjareeglitele ei ole liitsõnas osasõnade piirid enamasti markeeritud. Liitsõnade analüüsimine on eesti keele puhul üks keerulisemaid probleeme, kuna üheltpoolt on liitsõnamoodustus suhteliselt vaba, semantiliselt mõistlike liitsõnade genereerimine ja äratundmine on arvutile aga väga raske.

Loodud sõnastike süsteemis on lahendus esialgu väga üldine: omavahel saab kombineerida kõiki nimisõnu, osa määrsõnu saab olla liitsõna eelkomponendiks ning omadussõnad ja verbide partitsiipidest tuletatud nimi- ja omadussõnad võivad olla liitsõna järelkomponendiks.

Nimisõnade liitumine on korraldatud minisõnastikega N-LIIT (liitumine ainsuse nimetavas, nt **tütar***laps*), G-LIIT (liitumine ainsuse omastavas, nt **võtme***kimp*) ja PLG-LIIT (liitumine mitmuse omastavas, nt **õite***aeg*). Kuna ainult liitsõna viimane komponent käändub ning käändsõnalised eelkomponendid on kas ainsuse nimetavas või omastavas käändes (erandid välja arvatud), siis tuli nimisõnade jätkusõnastike arvu kahekordistada. Esimene sõnastik, millele juurtesõnastikust viidatakse, on hargnemissõnastik, millest saab edasi

minna kas liitsõnamoodustusse või käänamisse. Joonis 5 illustreerib liitsõnade moodustamise protsessi.



Joonis 5. Liitsõna *tütarlapsed* moodustamine sõnastike süsteemis

Mõnda tüüpi sõnade puhul võib tüvi liitumisel ka lüheneda, nt *laulmine+tund=laulmistund* (mitte *laulmisetund**), *inimene+vihkaja=inimvihkaja* (võimalik ka *inimesevihkaja*), nii et reeglid ei saa tegelikult olla üldised, tervet muuttüüpi haaravad. Kuid igasugune muuttüüpidega mittekattuvate loendite sissetoomine sõnastike süsteemi on äärmiselt kulukas, kuna sõnastike struktuur on idee poolest üles ehitatud morfotaktilisele struktuurile.

Määrsõnade kuulumist liitsõnade koosseisu reguleerib minileksikon PREFIX. Kui määrsõna saab olla liitsõna eelkomponendiks, siis on tema järel tüvesõnastikus viit leksikonile PREFIX. Sellest leksikonist saab edasi minna verbituletiste leksikoni, kuna enamasti tavalise nimi- ega omadussõna ette määrsõna ei liitu.

Liitsõna järelkomponendiks võib olla verbist tuletatud käändsõna, kuid mitte "puhas" verb – seetõttu ei olnud muud väljapääsu kui verbituletiste jaoks omaette sõnastik teha. Selle tagajärjel on iga verbi tüvi nüüd kirjas kaks (astmevaheldusega või kirjapildis mitte kajastuva astmevaheldusega) või kolm korda (astmevaheldusliku verbid) – üks kord verbijuurte sõnastikus, teine (ja kolmas) kord verbituletiste sõnastikus.

Näiteid liitsõnade morfoloogilisest analüüsist ja sünteesist programmiga *lexc*

Praegune sõnastike võrk võimaldab genereerida ja ära tunda nii normaalseid kui naljakaid liitsõnu. Viimased on, tõsi küll, vormiliselt korrektsed.

lexc> random-surf (Neid sõnu ei genereeritud järjest, see on valik mitmest seansist)
ehtejutud
hõbejõud
veahobust
arvuvabasti
grupiööõiguseks
õnnetusabigrupinägu

```
kõrvanõrku
unihambasoojuseaastateta
lexc> lookup silmarõõmuks
silm+rõõm+S+Sg+Tr
silm+rõõm+uks+S+Sg+N
```

Nagu näha, võib liitsõna morfoloogiline analüüs olla mitmene, kui osasõnade piire saab paigutada mitmesse kohta.

3.3.6. Sõnastike koostamise probleemid

Kahetasemelise morfoloogiamudeli sõnastike süsteem tundub esmapilgul võimas: viitade abil saab sõnajuured, tuletusliited, morfoloogilised tunnused ja lõpud sobivalt ühendada ning sõnavormid ongi olemas. Praktilises töös tekib aga rida probleeme.

1. Kuna eesti keeles on palju muuttüüpe, tekib väga palju minisõnastikke ning sõnastike süsteemis on raske orienteeruda.
2. Sõnastike maht suureneb verbituletiste lisamisel plahvatuslikult – iga verbitüvi tuleb 2–3 korda sisse – üks kord sõnastikku "Tegusõna" ning 1–2 korda sõnastikku "Verbituletis", sõltuvalt sellest, kas sõna on astmevahelduslik. Põhjus on selles, et verbid käituvad morfotaktiliselt hoopis teisiti kui verbituletised.
3. Mudeliga põhimõtteliselt sobimatu on loendite sissetoomine sõnastikku, kuid derivatsiooni ja liitsõnamoodustuse puhul on see möödapääsmatu. Sõnatuletus on küll osaliselt produktiivne, kuid tihti sõltub liite sobivus konkreetsest sõnast. Iga loend tähendab tegelikult sama viitstruktuuri kordamist teise nime all ning suurendab muidugi ka sõnastiku mahtu.
4. Ei õnnestu järgida põhimõtet, et sõnastikes kirjeldatakse morfotaktika põhimõtted ja tüvevariantide vaheldusmallid ning reeglitega tüvevariantide omavahelised fonoloogilised seosed. Tüvelõpuvaheldusi ei saa üldiselt reeglitega käsitleda, sest need on kujunenud sõna individuaalseks omaduseks.
5. Morfotaktika reegleid väljendav sõnastik peaks olema selline, kus sõnavormi moodustamisel morfeemidest vastab igale morfeemile täpselt üks sõnastik, aga eesti keele puhul tuleb sageli morfeeme tükeldada, mis muudab sõnastikud raskesti loetavaks.

3.4. Sõnavormide analüüs ja süntees koostatud sõnastike ja reeglite ning programmide *lexc* ja *twolc* abil

Koostatud sõnastike süsteem ja reeglid võimaldavad juurtesõnastikesse sisestatud tüvede piires genereerida ja ära tunda eestikeelseid sõnavorme. XRCE-st uurimisotstarbeliseks kasutamiseks saadud programm *lexc (***lex***icon* ***c****ompiler)* (Karttunen 1993) teeb õige struktuuriga sõnastikefailist lõpliku teisendaja ning programm *twolc (***two-l***evel* ***c****ompiler)* (Karttunen, Beesley 1992) kompileerib kõik reeglitefailis sisalduvad regulaarsete avaldistena esitatud reeglid lõplikeks teisendajateks. Programmis *twolc* on tervet reeglite kogu võimalik testida käsuga *lex-test*. Sisendiks tuleb anda mingi sõnavormi sõnastikuesitus ning programm genereerib reeglite põhjal pindsümbolite stringi, mis langeb kokku korrektse sõnavormiga, kui reeglid töötavad õigesti. Niimoodi saab reegleid muidugi pisteliselt kontrollida. Otstarbekas on reeglite kogu testida pärast iga uue reegli lisamist, et kontrollida, kas see töötab ootuspäraselt. Testimisel on võimalik kasutada ka süvastringide testfaili. Lõpuks saab programmiga *lexc* sõnastik-teisendaja ja reegel-teisendajad ühendada ning tulemust testida.

Programmis *lexc* on järgmised testimisvõimalused:

• Sõnavormide juhuslik genereerimine pind-, süva- või paralleelselt mõlemas esituses (vastavalt käsud *random-surf; random-lex* ning *random*).

Näide süsteemi poolt juhuslikult genereeritud sõnavormidest

```
lexc> random-surf
käed            +
pessa           +
öeldavaid       +
kohaehe         ?
vanalt          +
pimeduse        +
ülejätnuiks     ?
eemaltõppijad   ?
nähtuta         +
läksin          +
```

Korrektsete sõnavormide järele kirjutasin "+" mitte kasutusel olevate sõnavormide järele "?" Selliseid vigu põhujustab tuletatud sõnade ja liitsõnade üleproduktsioon.

• Üksiksõnavormi morfoloogiline analüüs (*lookup* sõnavorm).

Näide üksiksõnavormi morfoloogilisest analüüsist
lexc> lookup pead
pea+S+Sg+P (algvorm *pea*, sõnaliik – substantiiv, ainsuse osastav)
pea+S+Pl+N (algvorm *pea*, sõnaliik – substantiiv, mitmuse nimetav)
pidama+V+Ind+Pr+Sg2 (algvorm *pidama*, sõnaliik – verb, kindel kõneviis, olevik, ainsuse 2. pööre)

• Üksiksõnavormi morfoloogiline süntees (*lookdown* algvorm+morfoloogiline info); Noomeni puhul on morfoloogiliseks infoks sõnaliik+arv+kääne, verbi pöördeliste vormide puhul sõnaliik+kõneviis+aeg(+tegumood)(+isik), käändeliste vormide puhul sõnaliik+infinitiivi liik(+kääne). Sulgudes olevad osad ei esine kõikide muutevormide puhul.

Näide morfoloogilisest sünteesist,
mis demonstreerib ühtlasi paralleelvormide genereerimist
lexc> lookdown kallis+A+Spr+Pl+El
kalleimaist
kalleimatest

• Kõikvõimalike sõnavormide genereerimine (*check-all*). Selle käsu väljundit saab salvestada ka faili. Kui aga sõnastike võrk on tsirkulaarne (nagu see liitsõnamoodustuse sissetoomisest saadik on), siis on kõikvõimalikke sõnavorme lõpmatu arv ja see variant pole kasutatav.

Nagu ülaltoodud kirjeldusest nähtub, võimaldavad programmid *twolc* ja *lexc* kahetasemelist grammatikat testida sõnavormikaupa. Kuna aga kahetasemelises mudelis on keelekirjeldus programmidest sõltumatu, siis korrektselt kokku pandud eesti keele sõnastike ja reeglite olemasolul peaks XEROXis väljatöötatud morfoloogial põhinevat tarkvara (speller, infootsija, keelemõistataja jm) saama automaatselt rakendada eestikeelsetele tekstidele.

4. Mudeli sobivusest eesti keelele

Hinnates kahetasemelise morfoloogiamudeli sobivust eesti keelele võib esile tuua mudeli järgmisi häid külgi (Uibo 1999: 55).

1. Kahetasemelisus on üldiselt kasulik, kuna sõnastikuesitus võib sisaldada informatsiooni, mida sõna häälikkoostisest välja ei loe:
   1.1. Eriliselt saab tähistada need foneemid, millel on tüvevahelduse tõttu pindesituses rohkem kui üks variant. Nimelt ei

sõltu tüvevahelduse tüüp tänapäeva eesti keeles sageli tüve foneemilisest kujust.

1.2. Sõnastikuesitusse saab kirjutada reeglites kasutust leidvaid morfofonoloogilisi tunnuseid ja morfeemipiire.

2. Reeglid ei ole järjestatud. Järjestatud reeglitekogu koostamisel peaks arvestama kõigi eelnevate reeglite mõju kontekstidele.
3. Reeglites on võimalik viidata suvalisel kaugusel asuvale kontekstile. Näiteks saab kontrollida tüve lõpus asuvat sümbolit, teadmata tüve silpide arvu.
4. Kui süva- ja pindsümboli vastavus esineb mitmes kontekstis, millel pole sisu ega vormi poolest midagi ühist, siis võib vastavad kontekstid loetleda ühe ja sama reegli paremal poolel. See säästab uute süvasümbolite kasutuselevõtust, mis üldjuhul toob kaasa seosetuid tähiseid. Nt on vastavus *S:0* võimalik nii laadivahelduslikku sõna nõrgas astmes kui *S*-lõpuliste sõnade lõpus. Esimesel juhul asub *S* vokaalide vahel, aga teisel juhul tüve lõpus.
5. Sõnastikud võimaldavad mugavalt käsitleda
   a) fonoloogiliselt põhjendamatuid tüvelõpumuutusi;
   b) morfotaktika reegleid;
   c) produktiivset sõnatuletust ja liitsõnamoodustust (osaliselt).

Puudustena võib märkida raskusi mitmesümboliliste segmentide teisendamisel reeglitega ning morfotaktika reeglitest sõltumatute loendite sobimatust sõnastike süsteemi. Kahetasemelisse mudelisse peaks lisama ka täiustuse selles osas, et sõnajuure morfoloogilist informatsiooni saaks kustutada, kui derivatsioon tekitab uue sõnaliigi (Uibo 2000). Praegu on probleem lahendatud tüvede dubleerimisega tuletiste sõnastikus, mis kasvatab sõnastiku mahtu ebasoovitavalt kiiresti.

5. Võimalikud arendused

Juursõnastiku mahtu on võimalik poolautomaatselt suurendada, kasutades EKI-s olemasolevaid sõnastikke ja programmimooduleid (vt veebilehekülge www.eki.ee/tarkvara). Lähitulevikus on plaanis kohandada olemasolevat sõnastike süsteemi Ü. Viksi morfoloogilisele klassifikatsioonile (Viks 1994), et uute kirjete lisamine oleks hõlpsam.

Kooskõlaline ja leksikaalselt piisav eesti keele morfoloogia kirjeldus oleks aluseks eesti keele morfoloogilisele analüüsile ja sünteesile. Samal ajal laieneks XRCE keeletarkvara, mis eeldab just kahetasemelist morfoloogiakirjeldust, eesti keelele.

6. Kokkuvõte

Loodud on eesti keele kahetasemeline kirjeldus, mis koosneb 45 reeglist ja umbes 200 sõnastikust. Seejuures on põhilised eesti keele morfoloogias esinevad nähtused ära kirjeldatud. Senised katsetused näitavad, et kahetasemeline morfoloogiamudel on kasutatav eestikeelsete lihtsõnade analüüsiks ja sünteesiks. Samal ajal ei sobi mudel praegusel kujul eriti hästi eesti keele sõnatuletuse ega liitsõnamoodustuse modelleerimiseks. Mudeli heaks küljeks on kindlasti selle väike arvutuskeerukus, mis tuleneb realisatsioonist väikeste lõplike automaatidena. Teiseltpoolt sunnib formalism mõnikord sõnavorme tükeldama selliselt, et keelekirjeldus muutub raskesti loetavaks ja sõnastike struktuur ei järgi enam morfeemipiire – ka üksikmorfeemid on vahel tükeldatud ebamäärasteks osadeks ning morfoloogiline info pole täpses vastavuses sõnastikus oleva sõnaosaga. Kahjuks pole hetkel eesti keele kahetasemelise morfoloogia formalismil põhinevat morfoloogilist analüsaatorit veel võimalik teisi formalisme kasutavate morfoloogiliste analüsaatoritega võrrelda, kuna sõnastikus sisalduv sõnavara ei kata reaalsete tekstide vajadusi.

Kirjandus

Erelt, T. 1997. Eesti ortograafia. Tallinn: Eesti Keele Sihtasutus.

Hint, M. 1978. Häälikutest sõnadeni: emakeele häälikusüsteem üldkeeleteaduslikul taustal. Tallinn: Valgus.

Hint, M. 1997. Eesti keele astmevahelduse ja prosoodiasüsteemi tüpoloogilised probleemid (Typological Problems of Estonian Grade Alternation and Prosodical System). Dissertation at the University of Helsinki, Dept. of Baltic–Finnic Languages. Tallinn/Helsinki.

Kaalep, H.-J. 1999. Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös. (Dissertationes philologiae estonicae Universitatis Tartuensis – 7). Tartu: Tartu Ülikooli Kirjastus.

Karlsson, F. 1974. Phonology, Morphology and Morphophonemics. Gothenburg Papers in Theoretical Linguistics. Göteborg.

Karttunen, L. 1993. Finite-State Lexicon Compiler. Technical Report. ISTL-NLTT-1993-04-02. April 1993. Xerox Palo Alto Research Centre. Palo Alto, California.

Karttunen, L., Beesley, K. 1992. Two-Level Rule Compiler. Technical Report. ISTL-92-2. October 1992. Xerox Palo Alto Research Centre. Palo Alto, California.

Koskenniemi, K. 1983. Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. University of Helsinki, Dept. of General Linguistics. Publications No. 11. Helsinki.

Koskenniemi, K. 1997. Representations and finite-state components in natural language. – Finite-state Language Processing. Toim E. Roche, Y. Schabes. Cambridge, London: The MIT Press. 99–116.

Kuusik, E. 1996. Eesti tüvemuutuste süsteemi modelleerimine. Magistritöö. Tallinn. TTÜ.

Kuusik, E., Viks, Ü. 1998. Reeglipõhine morfoloogiline süntees. – Arvutimaailm 1, 43–45, 63; 2, 19–21.

Roche, E., Schabes, Y. 1997. Introduction. – Finite-state Language Processing. (Language, Speech, and Communication Series). Toim E. Roche, Y. Schabes. Cambridge, London: The MIT Press. 1–66.

Sproat, R. 1992. Morphology and Computation. ACL-MIT Press Series in Natural Language Processing. Cambridge, London: The MIT Press.

Uibo, H. 1998. Kahetasemeline morfoloogiamudel ja eesti keel. – Keel ja Kirjandus 1, 13–21.

Uibo, H. 1999. Eesti keele sõnavormide arvutianalüüs ja -süntees kahetasemelist morfoloogiamudelit rakendades. Magistritöö. TÜ arvutiteaduse instituut. Tartu.

Uibo, H. 2000. On using the two-level model as the basis of morphological analysis and synthesis of estonian. – Proceedings of the 12th Nordic Conference of Computational Linguistics NODALIDA'99 (ilmumas).

Viitso, T.-R. 1990. Eesti keele kujunemine flekteerivaks keeleks. – Keel ja Kirjandus 8, 456–461; 9, 542–548.

Viks, Ü. 1979. Kuidas formaliseerida tüvemuutusi. – Keel ja Kirjandus 11, 671–677.

Viks, Ü. 1980. Klassifikatoorne morfoloogia. Verb. ETA Keele ja Kirjanduse Instituut. Tallinn: Valgus.

Viks, Ü. 1982. Klassifikatoorne morfoloogia. Noomen. ETA Keele ja Kirjanduse Instituut. Tallinn: Valgus.

Viks, Ü. 1988. Fonoloogiliste ja morfoloogiliste mallide seosed. – Arvutuslingvistika sektori aastaraamat 1988. Tallinn. 160–166.

Viks, Ü. 1992. Väike vormisõnastik. 1. Sissejuhatus & grammatika. 2. Sõnastik ja lisad. ETA Keele ja Kirjanduse Instituut. Tallinn.

Viks, Ü. 1994a. Klassifikatoorne morfoloogia. (Dissertationes philologiae estonicae Universitatis Tartuensis). Tartu.

Viks, Ü. 1994b. Eesti keele morfoloogiline analüsaator. Automaatanalüüsi võimalused ja võimatused. – Keel ja Kirjandus 3, 150–163.

http://www.lingsoft.fi – keeletarkvarafirma Lingsoft.

http://www.xrce.xerox.com/research/mltt/ – XEROXi Euroopa uurimiskeskus, mitmekeelse teooria ja tehnoloogia uurimisrühm (multilingual theory and technology).

http://www.eki.ee/tarkvara – Eesti Keele Instituudi vabatarkvara.

Eesti keele reeglipõhise morfoloogilise ühestamise probleemseid kohti

Tiina Puolakainen
Tartu Ülikool

Käesolevas artiklis käsitletakse kitsenduste grammatikal (Karlsson jt 1995) põhineva morfoloogilise ühestaja (Puolakainen 1996, 1998; Müürisep, Puolakainen 1996, 1997) probleemseid kohti ning võrreldakse neid käsitsi ühestamisel tekkinud probleemidega (Kaalep jt 2000). Antakse lühiülevaade ka kitsenduste grammatika ühestaja tööpõhimõttest.

1. Sissejuhatus

Morfoloogiline analüsaator leiab sõnavormile kõik võimalikud morfoloogilised analüüsid, kuid ainult antud üksikut sõnavormi vaadates ei ole tal võimalik nende hulgast valikut teha. Enamuses rakendustes oleks aga hea teada mitte ainult seda, millised võimalikud tõlgendused on sõnal, vaid milline on tema tõlgendus konkreetses kontekstis. Näiteks sagedussõnastiku koostamisel on vaja teada, kas sõnavormi *peeti* lugeda nimisõnaks või verbiks. Selliste sõnavormide näiteid võiks tuua veelgi: *viis, sai, või.* Ilma iga sõnavormi morfoloogilist tõlgendust teadmata ei ole mõeldav ka lause (teksti) edaspidine (süntaktiline, semantiline) analüüs, ei saa hästi funktsioneerida infootsisüsteemid, masintõlke- ja dialoogsüsteemid. Ühestaja (*tagger*) ülesanne ongi sõna konteksti arvestades otsustada, milline kõikidest morfoloogilise analüsaatori pakutud analüüsidest õige on.

Erinevates keeltes on morfoloogiliselt mitmeti tõlgendatavate sõnade (sõnavormide) osakaal erinev: nt inglise keeles on mitme tõlgendusega 40% sõnadest, rootsi keeles üle 60% sõnadest, aga soome keele sõnadest võib ainult 11% leida mitu leksikaalset tõlgendust. Katsed näitavad, et eesti keeles on ligikaudu 40% sõnavormidest ilukirjandustekstides mitmeti tõlgendatavad (peaaegu 50% tõlgendustest, mida pakub morfoloogiline analüsaator, osutuvad konteksti mittesobivateks). Alternatiivsete analüüside tekkimise peamised põhjused on vormide homonüümia (nt sõnavorm *ilma* võib olla nimisõna genitiivis, partitiivis või aditiivis,

kaassõna või määrsõna) ja kategoriaalne mitmesus (nt kõikide partitsiipide puhul on vaja valida omadussõna ja verbi tõlgenduse vahel).

2. Lühidalt kitsenduste grammatika formalismist

Kitsenduste grammatika (*Constraint Grammar*) formalismi põhijooned esitas 1990 esmakordselt F. Karlsson Helsingi Ülikoolist. Edaspidi on selle väljaarendamisega tegeldud Helsingi Ülikooli üldkeeleteaduse osakonna arvutuslingvistika uurimisrühmas (Karlsson jt 1995). Inglise keele kitsenduste grammatika (ENGCG, Voutilainen jt 1992) on senini kõige täielikum konkreetse keele kitsenduste grammatika. Arendamisel on kitsenduste grammatikad soome, rootsi, taani, türgi, baski jpt keelte jaoks.

Kitsenduste grammatikas on analüüsi põhietappideks eeltöötlus, morfoloogiline analüüs, morfoloogiline ühestamine ja ka pindmine (*shallow*) süntaktiline analüüs. See on terviksüsteem, kus erinevad etapid on omavahel tihedalt seotud. Sõnastikku paigutatava informatsiooni valikul arvestatakse morfoloogilisele analüüsile järgnevate etappide (morfoloogilise ühestamise ja süntaktilise analüüsi) vajadusi. Grammatilisi tunnuseid esitatakse sõnadele lisatavate märgendite abil. Märgendid näitavad sõnaliiki, morfoloogilist infot (arv, kääne, pööre jne), süntaktilise analüüsi etapil ka süntaktilist funktsiooni. Sõna morfoloogilise tõlgenduse ja süntaktilise funktsiooni määramisel toimib sama põhimõte: alguses lisatakse sõnadele kõik võimalikud variandid ja siis eemaldatakse konteksti mittesobivad tõlgendused või märgendid spetsiaalsete reeglite ehk kitsenduste rakendamise teel.

3. Ühestaja tööpõhimõtted

Kitsenduste grammatika ühestaja vaatab korraga ühte lauset. Esiteks püüab ta lauses määrata osalausete piirid. Selleks rakendatakse spetsiaalseid osalausete piiride määramise reegleid. Need võivad olla näiteks järgmist tüüpi: lisa sõna tõlgendustele osalause piiri märgend, kui sõna kuulub hulka *et, sest, miks*. Või teine reegel: lisa sidesõnale osalause piiri märgend, kui talle eelneb koma ja mõlemal pool sidesõnas leidub verbi finiitne vorm. Osalause piire peab ühestaja teadma sellepärast, et ühestamise reeglite rakendamisel ei tohi tavaliselt reegli tingimusi kontrollides väljuda osalause piiridest, muidu ei pruugi reegli käitumine olla korrektne.

Järgmise etapina rakendatakse igale lause mitmesele sõnavormile ühestamise reegleid ehk kitsendusi, vaadeldes sõnu järjest lause algusest kuni lõpuni. Iga mitmese sõna korral püütakse rakendada reegleid, mis esitavad hulka tingimusi sõna konteksti kohta. Viimast tõlgendust ei eemaldata kunagi. Kui kõik reeglis esitatud tingimused sõna enda, tema kõrval olevate sõnade ning muude lauses kaugemal olevate sõnade kohta on täidetud, siis sõnal kustutatakse mõni tõlgendus (või valitakse välja üks ja kustutatakse ülejäänud). Reegel võib olla näiteks järgmist tüüpi: eemalda kaassõna märgend, kui eelmisel ega järgmisel sõnal ei ole nimisõna tõlgendusi. Tingimusi võib esitada ka kaugemate sõnade kohta, nt: loe finiitse verbi ainsuse tõlgendus õigeks ja kustuta kõik ülejäänud, kui lauses eespool leidub nimisõna ainsuse nominatiivis. Seda reeglit võib muuta ka keerulisemaks: loe finiitse verbi ainsuse tõlgendus õigeks ja kustuta kõik ülejäänud, kui eespool leidub ühene nimisõna ainsuse nominatiivis, sellest eespool pole komasid ega sidesõnu, ning kontrolli neid tingimusi ainult osalause piires.

Et tänu sellisele esimesele ühestamise ringile on mõned mitmesed sõnavormid saanud ühese tõlgenduse, siis on lootust, et uuesti lause algusest alustades on võimalik veel mõningaid reegleid rakendada: paljud reeglid nõuavad, et kontrollitav sõna oleks ühene – ainult siis võib teise sõna kohta järeldusi teha. Nii korrataksegi osalausete piiride määramist ja ühestamist vaikimisi veel kaks korda.

Eesti keele morfoloogilise ühestamise reegleid on praegusel hetkel üle tuhande. Nad teevad keskmiselt 2% vigu ja jätavad mitmeseks 10–15% sõnu. Edaspidi on artiklis juttu nende kahe protsendiarvu taga peituvatest lingvistilise kirjelduse formaliseerimise probleemidest.

4. Lingvistilised probleemid

Automaatsel morfoloogilisel ühestamisel ilmnevad lingvistilised probleemid kattuvad suures osas käsitsi ühestamisel tekkivate probleemidega. Kuna viimaseid on väga põhjalikult kirjeldatud artiklis (Kaalep jt 2000), siis ühiseid probleeme käsitleme järgnevas väga lühidalt ning kasutame võrdlemise hõlbustamiseks sama alajaotust.

4.1. Kaassõnad

Kaassõnade klassi avatus iseenesest ei ole automaatse ühestamise jaoks probleem, kui morfoloogiline analüsaator pakub kaassõna tõlgendust ning see kaassõna on kantud ka lisasõnastikku koos noomeni nõutava käände infoga. Probleeme tekib aga nende kaassõnadega, mis teises kontekstis võivad olla nimisõnad või määrsõnad või mõlemad.

Kaassõna reeglid on küllalt edukad, kuid juhuslike kokkusattumuste puhul tekib ka nendega vigu. Näiteks kaassõna *mööda* on nii pre- kui ka postpositsioon, mistõttu järgnevas lauses määratakse sõna *kooli* käändeks valesti partitiiv (tegelikult on siin kaassõnafraasiks *võimalust mööda*):

> Talviti pärast sõda sõitsime võimalust mööda _kooli_ ja koju talumeeste regede päradel.

Järgmises lauses määratakse sõna *kohale* kaassõnaks, kuigi ta seda ei ole:

> "Ma pean vist valvearsti _kohale_ kutsuma, kui sa mõistlikku juttu ei kuula," ähvardas ajakirjanik.

Suuri raskusi tekitab sõnarühm *kätte, käes, käest,* samuti kõik muud juhud, kus otsustamise kriteeriumid on semantilise iseloomuga.

Sõnadega *pähe, peas, peast* ning *koju, kodus, kodust* automaatsel ühestamisel probleemi ei teki tänu sellele, et nende sõnaliigi kuuluvus on üheselt ära otsustatud. Samuti ei teki probleeme ka sõnadega *kombel, moel, viisil* ning muude kaassõnadega, mis esinevad koos omastavas käändes noomeniga.

4.2. Asesõnad

Asesõnade semantilis-funktsionaalsesse rühma kuuluvust ühestaja otsustada ei püüa – kõiki ainult asesõna liigi poolest erinevaid tõlgendusi loetakse üheks, küll aga valmistavad probleeme sellised konkreetsed sõnad nagu *oma, üks* ja *teine*.

Sõnavormil *oma* on formaalselt võimalik vahet teha määrsõna tõlgenduse ja teiste tõlgenduste vahel. Verbi tõlgendus allub üldistele reeglitele. Asesõna ja omadussõna vahel aga ei saa selget piiri tõmmata: mõlemad laiendavad nimisõna ja väljendavad kuuluvust (Viks 1972), nt:

Peetril on mure oma töö pärast.	(asesõna)
Igaühel on oma maailm, omad mured ja rõõmud.	(omadussõna)
Oma sõpru ja vaenlasi ei unustata niipea.	(asesõna)
Oma võlu ja veetlus peitub selleski töös.	(omadussõna)

Mõned üksikud juhud on siiski välja eraldatud 11 reegliga, mis on koostatud põhiliselt Ülle Viksi töödes (Viks 1972, 1980) toodud näidete abil.

- *oma* on pronoomen lause lõpus; juhul, kui talle vahetult eelneb substantiivne pronoomen *(ta, ma* jne) genitiivis; juhul, kui järgmine on verb, määrsõna või kaassõna. *oma* ei ole omadussõna, kui järgmine sõnavorm ei ole nimetavas, omastavas või osastavas käändes. Kuigi see reegel põhineb tõsiasjal, et omadussõna peab ühilduma oma põhisõnaga, võib see siiski põhjustada vigu, kuna *oma* ühildumine ei ole nii üldine nagu teistel adjektiividel (Viks 1972).
- *oma* on määrsõna, kui vahetult järgmine on põhiarvsõna, hulgasõna või ajamääratlus.

Sõnavormi *üks* (ja ka kõigi selle sõna käändevormide) esinemisel tekstis tuleb valida arvsõna või asesõna tõlgenduste vahel (mõnedel käändevormidel lisanduvad veel nimisõna *ühe* vormid). Üldjuhul ei ole võimalik neid tõlgendusi eristada, nt: *Üks* (arvsõna) *tema laps õpib ülikoolis. Ta tutvus ühe* (asesõna) *ameeriklasega.* Mõned erijuhtumid on siiski püütud välja eraldada.

- *üks* on asesõna, kui järel on *teine* (*Tulid üks ees, teine järel. Üks häda teise otsa.*).
- *üks* on arvsõna, kui järgmine on finiitne verb.
- *üks* on arvsõna, kui talle eelnevad sõnad *ainult, veel, kõigest, vaid* (*ainult üks päev*).
- *üks* on arvsõna, kui talle järgneb nimisõna, omadussõna, asesõna või järgarvsõna mitmuse osastava või seestütleva vormis.
- *üks* on arvsõna, kui talle järgneb nimisõna, omadussõna, asesõna või järgarvsõna mitmuse genitiivis (genitiivse laiendiga) ja seejärel on sõna *hulgast, seast* või *keskelt.*
- *üks* on arvsõna, kui järgneb lause või osalause piir (*Lapsi on tal üks. Nende hulgas on üks, kes teeb pahandusi.*).

Sõna *teine* sõnaliigi otsustamisel kasutatakse peamiselt kaalutlust, et kui lauses eespool esines *üks,* siis on nad mõlemad asesõnad.

Erinevalt käsitsi ühestamisest valmistab automaatsel morfoloogilisel ühestamisel raskusi asesõnade *ta/tema, me/meie* ja *te/teie* nimetava ja omastava käände eristamine. Nt lauses *Ta tahtis järele katsuda tunnet, mis öösel tema hingenurka oli asunud* saab inimene aru, et *tunne oli asunud tema hingenurka*, ilma tähendust teadmata jääb aga ka selline tõlgendamise võimalus: *millisel öösel oli tema asunud kuhugi nurka.* Seevastu aga sellisest eraldiseisvast lausest nagu *Tema surivoodi juures näiteks* ei saa ka inimene aru, kas oluline on *tema* kellegi teise surivoodi juures või siis koht (*tema surivoodi juures*).

Alati pole võimalik ka sõnavormide *kes* ja *mis* puhul määrata, kas nad on ainsuses või mitmuses. Põhiliseks võimaluseks seda lahendada on:

a) järgneva nimisõna arvu järgi, kui see on nimetavas käändes;
b) osalauses oleva verbi arvu järgi;
c) kui tegemist on sõnavormiga *kes* või *mis* algava kõrvallausega, siis eelmise osalause viimase sõna arvu järgi. Mõned juhtumid jäävad ka siin lahendamata, nt: *Siis saab varakult teada, mis on mis ja kuidas see käib.* Siin on verbivormi *on* arv määramata ja ka arvukategooria suhtes üheseid nimisõnu osalauses (*mis on mis*) ei ole.

4.3. Määrsõnad ja sidesõnad

Sõnade *aga, nagu, kui* sõnaliigi määramisega tekib probleeme põhiliselt just seetõttu, et inimene määrab selle sõna semantilise funktsiooni alusel, mida morfoloogilisest infost ei ole võimalik tuletada. Sõna *kui* kohta otsustab ühestaja siiski enamasti õigesti. Kui ta alustab kõrvallauset, siis eraldi reegel püüab eraldada need juhud, kus *kui* väljendab määra astet, ulatust ning on seega määrsõna, nt *Nüüd alles jõudis tema teadvusse, kui suure asja ta on ette võtnud.* Siin on põhiliselt kasutatud asjaolu, et adverbile *kui* peab järgnema kas adverb või omadussõna, ja lisaks hulk juhtumeid, kus *kui* on kindlalt sidesõna. Kui aga sõnavormile *kui* tõesti järgneb adverb või omadussõna, siis ei ole võimalik määrata, kas ta on side- või määrsõna, nt:

Kui homme sajab, jätan minemata. (sidesõna)
Kui kaua see kestab, ei tea keegi. (määrsõna)

Vaatleme prooviks sellist reeglit: valida määrsõna tõlgendus, kui ühestatavale sõnavormile järgneb üks sõnadest: *mitu, vähe, palju, hea, halb, hästi, kaua, vana, noor, raske, kerge.* Ka sellise piiratud hulgaga ei anna reegel alati õigeid tulemusi – vigu tekib juhul, kui järgmine omadussõna kuulub küll loetletud hulka, kuid ei kuulu siiski *kui*ga kokku, nt: *See ikka ei ole tõsi, kui hästi mõelda.*

Sõnavormide *aga, või* ja *kuid* puhul on valida sidesõna- ja määrsõna tõlgenduste vahel. Sõnavormi *või* puhul lisanduvad veel nimisõna kaks käänet – nimetav ja omastav – ning verbi vorm (*võima*). Põhireegel sidesõna ja määrsõna tõlgenduste eristamiseks on see, et kui eelmine ja järgmine sõna on samast sõnaliigist ja samas vormis, siis on *aga, või* ja *kuid* sidesõnad. Erinevus on selles, et *või* ees ei pruugi koma olla, *kuid* ja *aga* ees seevastu kindlasti on.

4.4. Verbiga seotud ühestamisprobleemid

Sõnavormi *on* puhul on probleem ainsuse või mitmuse vormi valimises (ka põhi- või abiverbi tõlgenduse valimises, aga see on tihedalt seotud partitsiipide probleemiga). Koostatud reeglites püütakse arvu määrata oletatava aluse põhjal ning arvestades konjunktsiooni: kui *on* järel leiduvad kaks rinnastatud nimisõna või asesõna ainsuse nimetava vormis, siis *on* on mitmuses (lisatingimuseks on see, et osalauses edasi ei tohi olla nimetavat; see välistab vigu, nt lauses: *Suurepäraselt on edasi antud võitja ja kaotaja eetos.*).

Selle reegli puhul aga tekib ka vigu, nt sellises lauses: *Loomulikult tema õueskulptori Lysippose tehtud originaalist, sest tolle materjaliks on kuld ja elevandiluu.*

Kõiki juhtumeid need reeglid ei lahenda, nt: *Siis saab varakult teada, mis on mis ja kuidas see käib.* Siin jääb verbivormi *on* arv määramata, kuna ühese arvuga nimisõnu osalauses (*mis on mis*) ei ole.

Leidub üsna palju huvitaval kujul verbi ja nimisõna vormide kokkulangevusi: *tuli, tuleks, viis, sai, või, peeti, oma, tee, too, päris, jälgi, korda, pead, pea.* Mõned neist lahendatakse üldiste reeglitega, aga teiste jaoks on kirjutatud eraldi reeglid. Nt sõnavormi *tuli* (ja ka *tuleks*) puhul eraldavad reeglid need juhtumid, kus see sõnavorm on väga suure tõenäosusega verb, nimelt siis, kui osalauses leidub ees või taga verbi *da*-tegevusnime vorm.

4.5. Verbide ja omadussõnade vahekorrast

Põhiliseks raskuseks on partitsiibid. Mineviku partitsiipide analüüsiks ei ole senini head lahendust leitud. Kasutatakse vaid kõige lihtsamaid reegleid, mis eemaldavad verbi tõlgenduse juhul, kui liitaja või eituse moodustamine ei ole võimalik, või eemaldavad adjektiivi tõlgenduse juhul, kui see sõna ei saa olla öeldistäiteks (lauses puudub verb *olema*) ning ei saa olla ka atribuudiks (lauses ei leidu nimisõna, mida laiendada). Kuid tihti on lauses formaalselt võimalik moodustada nii liitaega kui laiendada nimisõna (tänu sellele, et mineviku partitsiibid ei käändu, võivad nad laiendada suvalises käändes nimisõna).

Oleviku partitsiipidega tekib probleeme vaid seoses sellega, et nende mitmuse nimetava käände vorm langeb kokku vastava verbi oleviku 3. pöörde mitmuse vormiga. Nt lauses *Kas siin on alati olnud ridamisi neid räämas üheksateistkümnenda sajandi maju, palgid seinu toetamas, aknad papiga kinni löödud ja katused lainelise plekiga kaetud, lagunevad aiamüürid igasse külge vajumas?* ühestatakse sõna *lagunevad* põhiverbiks, mitte adjektiiviks.

Veel üheks verbi vormidega kokkupuutepunktiks on oleviku partitsiibi ainsuse partitiivi käände kokkulangemine verbi *vat*-vormiga. Nt lauses *Aga hoolimata tema heidutavast välimusest oli tema käitumises teatavat sarmi* jäetakse adjektiivina esinevale sõnale *teatavat* ka verbi tõlgendus alles. Analoogiline situatsioon on ka lauses *Ta tundis tema vastu sügavat huvi*: sõnale *sügavat* jäetakse mõlemad tõlgendused.

4.6. Omadussõnad

Omadussõnade morfoloogilisel ühestamisel kasutatakse üldjuhul adjektiivi–atribuudi ja põhisõna ühilduvuse reegleid ning konjunktsioonireegleid, mis lihtkonstruktsioonides valivad enamasti adjektiivide puhul õigeid tõlgendusi. Kuid on mitmeid erijuhtumeid, kus sellistest lihtsatest reeglitest ei piisa.

Käändumatute omadussõnade puhul esineb kõige rohkem probleeme adjektiivi ja määrsõna tõlgenduste vahel valimisel. Nt lauses *Partei liikmed ei tohtinud tavalistes poodides käia ("vabalt turult ostmas" nagu öeldi), aga sellest ei peetud rangelt kinni, ...* ei ole võimalik formaalselt otsustada, kas *vabalt* peaks olema määrsõna või nimisõnaga *turg* ühilduv omadussõna. Lauses *See oli tohutu*

kiiskavvalgest betoonist püramiidne ehitis, mis kerkis astanguliselt 300 meetri kõrgusele jäi sõna *tohutu* mitmeseks selle tõttu, et järgmisel sõnal ei olnud morfoloogilist tõlgendust (see ei sisaldunud morfoloogilise analüsaatori sõnastikus) ning mingeid otsuseid ei olnud võimalik teha. Lauses *Kas siin on alati olnud ridamisi neid räämas üheksateistkümnenda sajandi maju, palgid seinu toetamas,...* on sõnal *räämas* adjektiivi, määrsõna ning ainsuse sisseütlevas käändes nimisõna tõlgendused, millest ühestaja reeglid ei saa õiget välja valida.

Kui morfoloogiline analüsaator ei anna substantiivi rollis kasutatavale omadussõnale nimisõna tõlgendust (nt ei anta seda kõikide *nud-, tud*-partitsiipide puhul), ei ole kitsenduste grammatikaga see probleem kuidagi lahendatav (reeglid saavad ainult eemaldada, mitte lisada tõlgendusi, igale adjektiivile substantiivi tõlgenduse lisamine põhjustaks rohkem mitmesust kui lahendaks). Ja juhul, kui antakse mõlemad tõlgendused (*valge, must, blond* on enamasti omadussõnad, kuid neid kasutatakse tihti nimisõna rollis; *haige, vana, kaotaja, võitja* on põhiliselt nimisõnad, kuid esinevad ka omadussõna rollis), ei ole alati lihtne valida õiget varianti. Lauses *Mõeldav oli seegi, et jälgiti kogu aeg kõiki* ei ole ühestaja võimeline otsustama, kas sõna *mõeldav* on adjektiiv või substantiiv (sellel konkreetsel juhul on see öeldistäide ja järelikult adjektiiv). Kuna mõnikord oleks morfoloogilise tõlgenduse valikul aidanud süntaktilise funktsiooni tundmine, siis on plaanis proovida koostada reegleid, mis kasutavad ka süntaktilisi funktsioone, ja rakendada neid pärast süntaktiliste funktsioonide määramist.

4.7 Ühend- ja väljendverbid

Probleemiks on asjaolu, et paljud ühendverbi koosseisus olevad adverbid (afiksaaladverbid) võivad teatud kontekstis olla ka kaassõnad (*üle – üle elama* ja *üle jõe*) või nimisõnad (*välja – välja minema* ja *väli*). Ühel sõnavormil saavad esineda ka kõik kolm võimalust, nt sõnavorm *kätte* võib olla afiksaaladverb (*kätte maksta*), kaassõna (*Peetri kätte*) või nimisõna aditiivis (*panen kindad kätte*). Tegelikult on probleem veelgi laiem, kuna adverbi võimalus on ka väljaspool ühendverbi. Nt sõnavormid *vahel, kõrvalt* võivad olla iseseisvad adverbid, kaassõnad ja ka nimisõnad. Kuid *pealt* võib olla nii afiksaaladverb kui ka iseseisev adverb (*pealt vaatama* ja *Võta pealt!*), samuti kaassõna ja ka nimisõna.

Reeglites on kajastatud ainult mõned sellised ühendverbid, tegelikult on neid palju rohkem. Kõiki neid reeglite kujul esitada on aga väga kohmakas ja ei paista olevat just elegantne lahendus. Paremaks lahenduseks oleks hiljem paigutada selline informatsioon (iga verbiga seostuvad afiksaaladverbid) kas otse sõnastikku või preprotsessorisse. Siiski võib tulla ka probleeme: see, et osalauses esineb teatud verbi vorm, ei tähenda alati, et nad koos võimaliku afiksaaladverbiga moodustavad ühendverbi. Nt kaks lauset: *Ma elasin üle vapustuse* ja *Ma elasin üle tänava.* Esimesel juhul on tegemist ühendverbiga (*üle elama*), teises lauses aga kaassõnaga.

Sõna *mööda* puhul tekivad samad probleemid – *mööda minema* ei pruugi olla ühendverb, nt: *Siin kavatses ta paremale pöörata, minna mööda järgmist tänavat kuni leivapoeni ja siis pöörata vasakule.*

4.8. Nimisõna käände valik

Suuri raskusi õige käände valikuga tekib selliste sõnade puhul, millel langevad kokku nimetava, omastava ja osastava käände vormid (*maja, loetelu, teadusharu, asjaolu, koostisosa, kala*); omastava, osastava ja lühikese sisseütleva (aditiivi) käände vormid (*jaama, pealkirja, konverentsi, probleemi, inimkultuuri*); nimetava ja omastava käände vormid (*kodumaa, uurimistöö, aasta, toodetu, jutustaja*).

Nimisõna käändeid püütakse valida konjunktsioonireeglitega, arvestades eelneva omadussõna käänet, eelneva või järgneva kaassõna või arvsõna käänet ning asjaolu, et lauses peab olema alus ja enamasti ka sihitis (kuigi need võivad olla nii nimetavas kui omastavas ja osastavas käändes).

Üheks raskemaks probleemiks on esimese kolme käände ühestamine: nominatiivi ja genitiivi tõlgendused jäävad tihti mõlemad alles. Formaalselt pole selge, miks näiteks ei või järgmises lauses *tema* olla subjekt: *Tema[nom, gen]*[1] *põiklevatest vastustest võis fantaasiat appi võttes välja lugeda, et abahakiiridel kehtis midagi ürgkondliku kommuuni või patriarhaadi taolist.*

Formaalselt on raske määrata ka seda, miks ei saa *Antti* olla genitiivis ja õppetöö nominatiivis (vrdl *Antti õppetöö algusest peale ei edenenud*): *Tema veidrused olid tuntud ning seepärast ei teinud*

[1] Järgnevates näidetes on alla joonitud õige variant.

erilist muret asjaolu, et Antti[nom, gen] õppetöö[nom, gen] alguseks tagasi ei jõudnud.

Nüüd polnud ta seda teinud ning oli asetanud kateedrijuhataja[nom, gen] täbarasse olukorda. Siin puudub teises osalauses formaalne subjekt ja ühestaja arvab, et *kateedrijuhataja* on nominatiivis.

Seda peasilitamist sai vahel isegi liiga[nom, part, määrsõna]. Määrsõna asemel valitakse hoopis nimisõna ainsuse nominatiivi tõlgendus.

Raske on otsustada, et esimeses lauses on *raha* partitiivis, kui teises analoogilises lauses on *poiss* nominatiivis:

Katsugu raha[nom, gen, part] paremini hoida.
Katsugu poiss[nom] paremini õppida.

Järgmises lauses on sõnavormil *süüd* huvitaval kombel kolm tõlgendust – *süü* mitmuse nominatiivis ning ainsuse partitiivis ja *süüd* ainsuse nominatiivis: *Süüd[nom, part] oli siin muidugi ka naisel.* Ühestaja eelistab sellises lauses aluse positsioonis ainsuse nominatiivi tõlgendust ning seega ka eksib.

4.9. Muud

Esineb ka põhimõtteliselt lahendamatuid probleeme. Sõnavormile *sool* antakse kolm ühesugust nimisõna ainsuse nominatiivi võimalust ja valida neist pole võimalik. *Näis kuulduvat* – kas on see *kuulma* või *kuulduma*? Kuidas saaks otsustada, et lauses *Trammis[omadussõna, määrsõna, nimisõna ines] oli palju inimesi.* on *trammis* nimisõna, aga mitte määrsõna, kui on olemas ka selline lause: *Eile[määrsõna] oli palju inimesi*?

Formaalselt võiks ka järgmises lauses olla *hukas* predikatiiv (omadussõna):

Noorus olevat alati hukas[omadussõna, määrsõna] olnud.
Noorus olevat alati ettevaatamatu[omadussõna] olnud.
Miina tütar läks linna[gen, part, adit] teenijaks. – Võibolla on olemas linna teenija?

Ka eelnevast sõnast ei ole alati kasu käände määramisel:

Teistele leiame muud[sg part, pl nom] tööd[sg part, pl nom].
Teistele leiame muud[sg part, pl nom] riided[pl nom].

Peale erinevate vormide korrapäraste kokkulangemiste esineb ka juhuslikke kokkusattumusi. Nt lauses *Kusagil oli ta siiamaani elus*

ja haudus oma salaplaane ei ole ilma semantikata võimalik aru saada, et *elus* peaks olema kindlasti adjektiiv, mitte substantiiv *elu* ainsuse seesütlevas käändes, eriti kui samas osalauses esineb ka kohamäärsõna *kusagil*.

5. Kokkuvõtteks

Automaatsel reeglipõhisel morfoloogilisel ühestamisel ilmnevate lingvistiliste probleemide ring kattub osaliselt käsitsi ühestamisel tekkivate probleemide ringiga, kuid ei lange sellega päris kokku. Jämedalt võttes võib morfoloogilise ühestamise probleemseid kohti jaotada kolme rühma. Esiteks sellised, mis käsitsi ühestamise etapil on põhjalikult läbi arutatud ja vastuvõetud otsused on niivõrd detailsed ja konkreetsed, et on suurepäraselt formaliseeritavad ning automaatse ühestamise puhul probleeme enam ei tekita. Teise rühma kuuluvad sellised probleemid, mis – kuigi on käsitsi ühestamisel arutlusel olnud ja vastavaid otsuseid ka vastu võetud – on siiski automaatse ühestamise jaoks tarbetud, kuna need otsused põhinevad lause ja konkreetsete sõnade tähendusel (semantikal). Ning lõpuks, kolmandasse rühma võib eraldada selliseid probleeme, mida inimene lahendab enda jaoks täiesti märkamatult, formaalselt aga pole neid kuigi kerge käsitleda. Kahe viimase rühma puhul on võimalikud järgmised lahendusteed.

1. Lihtsustada morfoloogilise kirjelduse detailsust. Nt mineviku partitsiipide puhul mitte eristada omadussõna ning verbi tõlgendust, vaid märgistada neid lihtsalt partitsiipideks (mõnede keelte puhul on nii ka tehtud). Selliste lähenemiste kasutuselevõttu tuleb muidugi eraldi kaaluda sõltuvalt laiemast ülesandest (kas morfoloogilist ühestamist tehakse infootsimise hõlbustamiseks, süntaksianalüüsiks, tõlkeabi programmis jne).

2. Proovida morfoloogilisel ühestamisel kasutada süntaktilise analüsaatori poolt leitud süntaktiliste funktsioonide märgendeid, s.t teha ühestamist täiendavalt pärast süntaktilist analüüsi.

3. Kasutada allesjäänud mitmesuste lahendamiseks statistilisi meetodeid.

Kirjandus

Kaalep, H.-J., Muischnek, K., Müürisep, K., Rääbis, A., Habicht, K. 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? – Keel ja Kirjandus, ilmumas.

Karlsson, F. Voutilainen, A., Heikkilä, J., Anttila, A. 1995. Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text. Berlin, New York: Mouton de Gruyter.

Müürisep, K., Puolakainen, T. 1996. Eesti keele automaatne süntaktiline analüüs kitsenduste grammatika abil. – A&A 10, 25–27; 11, 23–28.

Müürisep, K., Puolakainen, T. 1997 Eesti keele automaatne süntaktiline analüüs kitsenduste grammatika abil. – A&A 1, 47–53; 2, 37–39; 3, 30–33; 4, 27–31.

Puolakainen, T. 1996. Eesti keele morfoloogiline ühestamine kitsenduste grammatika abil. – Tartu Ülikool, Arvutiteaduse Instituut. Magistritöö.

Puolakainen, T. 1998. Eesti keele kitsenduste grammatika morfoloogiline ühestaja. – Keel ja Kirjandus 1, 37–46.

Viks, Ü. 1972. Sõnast "oma" eesti keeles. – Keel ja struktuur 6, Tartu. 129–138.

Viks, Ü. 1980. Mõne pronoomeni kasutamisest. – Keel ja Kirjandus 10, 617–619.

Voutilainen, A., Heikkilä, J., Anttila, A. 1992. Constraint Grammar of English: A Perfomance-Oriented Introduction. Publication no 21. Department of General Linguistics, University of Helsinki.

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 tekstikorpusi ehk -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 mittestandardne, 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 lihtsalt 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 takistab süvenemist.

Käsud ("klotsid"), millest lingvistile vajalikke filtreid koostatakse, on järgmised.

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 teksti, 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" 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-indeksite esitamiseks. On hea, kui neid saab kasutada samasuguste ehitusklotsidena nagu UNIXi enda käske.

Eesti keele puhul on vaieldamatult vajalikuks ehitusklotsiks morfoloogiline analüsaator, st 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 järgmine: tekst → lausepiiride leidmine → morfoloogiline 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 ehk 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 legoklotsiks, 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, st 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 (detailset kirjeldust vt Kaalep 1997, 1998).

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.

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 teatmematerjalide 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 *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 inglise-keelsete 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 häälдust 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. Nt 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: 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, *Hidden Markov Model*, HMM; vt Kaalep, Vaino 1998) nime all tuntud tõenäosuslik mudel. 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-i, siis konkreetsele lausele võib vastata mitu võimalikku M-de järjestust, aga ainult üks neist on õige.
3. Mõned järjestused on antud keeles tüüpilised, mõned mitte.
4. Võimalikest järjestustest tuleb valida kõige tüüpilisem, so. kõige tõenäolisem. See ongi antud lause puhul õige.
5. Uue lause M-de 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 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%). 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 umbes 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 versiooni (http://www.filosoft.ee/html_morf_et/morfoutinfo.html) kasutab ka meie poolt kirjeldatud 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 ning Puolakainen 2000. Siinkohal piisab tõdemusest, et teoreetilistes käsitlustes nagu Valgma, Remmel 1970 ja EKG 1995 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.

Kirjandus

EKG I 1995 = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. Eesti keele grammatika. I. Morfoloogia. Sõnamoodustus. Tallinn: ETA Eesti Keele Instituut.

Kaalep, H.-J. 1997. An estonian morphological analyser and the impact of a corpus on its development. – Computers and Humanities. 31, 115–133.

Kaalep, H.-J. 1998. Tekstikorpuse abil loodud eesti keele morfoloogia-analüsaator. – Keel ja Kirjandus 1, 22–29.

Kaalep, H.-J., Vaino, T. 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. – Keel ja Kirjandus 1, 30–38.

Kaalep, H.-J., Muischnek, K., Müürisep, K., Rääbis, A., Habicht, K. 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? – 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.

Puolakainen, T. 1998. Eesti keele kitsenduste grammatika morfoloogiline ühestaja. – Keel ja Kirjandus 1, 37–46.

Valgma, J., Remmel, N. 1970. Eesti keele grammatika. Tallinn: Valgus.

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.

Leksikaalse info kodeerimine

Margit Langemets
Eesti Keele Instituut

Olles juba mõnda aega uurinud sõnaraamatute "elu" arvutis meil ja mujal maailmas, püüab käesoleva loo autor mõne kandi pealt avada seda valdkonda ka teistele. Siinne sissevaade kõneleb üldisemalt arvutileksikograafiast kui omaette uurimisharust, mõningatest käibel olevatest terminitest ja mõistetest, lähemalt vaadeldakse *kuidas* sõnaraamatus olevat infot kirjeldada, piirdudes siingi üksnes valikuga. Ei vaadelda aga seda, *mis* infot sõnaraamat üldse sisaldab ega ka sõnaraamatute kasutamist keeletöötluses.

1. Taustaks

Arvutis olevate sõnakogude ajalugu on küll pea sama pikk kui arvutava masina enda eluiga (umbes 50 aastat), ent arvutilingvistide tõsist huvi õnnestus arvutisõnastikel äratada tunduvalt hiljem, alles 1980. aastatel. Tolleks ajaks olid seljataha jäänud esimesed tõsisemad masintõlkekatsetused ja edukad süntaksianalüüsi aastad. Mööda oli ka massiline traditsiooniliste sõnaraamatute arvutisse sisestamise laine, mille käigus väga paljud kultuurkeeled varustasid end enam või vähem mahukate leksikaalsete andmekogudega. Arvutilingvistidel tekkis vajadus ja võimalus avardada oma tegevusvälja sõnaraamatute suunas.

Arvutileksikograafia sündi dateeritakse aastaga 1986, mil Itaalias toimus kuulus Grosseto õpikoda (Walker jt 1995). Grossetos püüti leksikaalse uurimistöö ülesandeid määratleda just arvutilingvistilisest vaatepunktist ja sõnastati ka otsesed tegevusjuhised:

- uurida leksikaalset infot arvutisõnastikes, analüüsida sõnaartikli olemust ja struktuuri;
- luua ühtne metaformaat sõnaartikli kirjeldamiseks;
- uurida võimalusi kasutada ühte ja sama arvutisõnastikku nii inimkasutaja kui ka keeletöötlussüsteemi jaoks;
- julgustada ja veenda kirjastajaid oma sõnastikke uurijate käsutusse loovutama;
- (statistilisi meetodeid kasutades) koguda korpusest andmeid sõnade, sõna tähenduste ja kasutuse kohta;

- arendada tekstifailidest koosnevaid andmebaase ja nende töötlusvahendeid jne jne.

Sõnastike arvutis olemise arenguteel nähakse vähemalt kolme järjestikust astet: 1970. aastatel lihtsalt arvutisse sisestatud tekstidelt on liigutud 1980. aastate leksikaalsete andmebaasideni, sealt omakorda leksikaalsete teadmusbaasideni.

1970. aastate arvutisõnastikud olid enamasti arvutisse sisse tipitud nagu trükimasinasse (*machine readable dictionary,* MRD). Ainsaks sihiks oli need arvuti abil välja anda, mistõttu sõnastiku tekst pikiti täis kõikvõimalikke polügraafilisi, kirjastiili ja leheküljekujundust suunavaid sümboleid.

1980. aastate leksikaalne andmebaas (*lexical database,* LDB, ka *machine-tractable dictionary*, MTD) on ulatuslik, kogu keelt ammendada püüdev infokogu, mis varasemaga võrreldes tunduvalt enam arvestab sõnastiku kui struktureeritud objekti iseloomu. See on justkui sõnastiku arvutilingvistiline, märgatava struktuuriga mudel (Calzolari 1995), mida arvutiprogrammide abil hõlpsasti uurida ja kasutada saab. Sisu mõttes on see eelkõige, nagu nimigi ütleb, leksikaalse info kogu, mis tähendab, et kõigekülgselt on kirjeldatud sõnu, peaasjalikult sõnade vormi ja funktsiooni, mingil määral ka tähendust.

Ambitsioonikate leksikaalsete andmebaaside kõrvale tekkisid 1980. aastatel ka nn ühe-sõnaraamatu-andmebaasid, mis esitavad ainult ühe sõnastiku põhjalikult analüüsitud materjali. Pioneeriks oli ses vallas R. Amsleri doktoritöö (Amsler 1980). Akadeemiline areng on sujuvalt suundunud ärisse – nüüdisajal ongi igal korralikumal kirjastusel omad, üksiksõnaraamatute alusvormile (*master source*) ehitatud üks- ja/või kakskeelsed leksikaalsed andmebaasid. Pole sõnaraamatukasutajat, kes poleks kuulnud Oxfordi, Longmani, Collinsi, Cobuildi, Langenscheidti või Websteri sõnaraamatutest.

1980. aastate kuulsaim arvutileksikograafiline hiigelprojekt teostati Kanadas, Waterloo ülikoolis, kus arvutisse sisestati *Oxford English Dictionary* 13 köidet ja 4 lisaköidet (kokku 500 tuhat sõnaartiklit, sh 1,8 miljonit näitetsitaati). Sõnastikust kujundati spetsiaalne struktureeritud tekstiandmebaas, tekst märgendati tollal uudse SGML-i abil ning tekstist info otsimiseks töötati välja oma päringusüsteem PAT. Projekt oli väga hästi rahastatud: 5 aastaga oli kogu tekst sisse tipitud (selleks kulus 120 inimaastat) ja korrektuur

loetud (60 inimaastat[1]) ning sõnaraamatu 2., täiendatud trükk nägi ilmavalgust juba 1989. aastal.

Arvuti kasutamise ühe suuna – keelandmete kogumise ja sorteerimise – ammendas 1980. aastate alguse Cobuildi sõnastikuprojekt: loodi 20 miljonist sõnast koosnev inglise tekstikorpus, sõnu sorditi kõikvõimalikesse konkordantsidesse ning saadud alusmaterjalile toetudes koostati Cobuildi sõnaraamat (Ide jt 1997).

1980. aastate lõpust alates kõneldakse ka leksikaalsetest teadmusbaasidest (vanem tarvitus: teadmiste baas; *lexical knowledge base,* LKB). Termini ristiisad on R. Amsler ja D. Walker, kes pidasid hädatarvilikuks kaasata loomuliku keele töötlusse tegelikkuse semantiline jaotus ja selle formaliseerimine. Olemasolevaid analüsaatoreid oli pessimistlikult kritiseeritud, sest neil puudus korralik sõnaraamatutugi, ka olid leksikaalsete mõistete esituses toimunud suured nihked: jõudsalt olid arenenud semantilised võrgustikud ja nende analüüsimehhanismid, kasutuses olid loogilised formalismid. Keele abil küll väljendatakse tähendust, ent mis on sõna tähendus? Kas see on sõnade kasutus keeles, nagu arvas Wittgenstein, või teaduslik teadmus asjade taga, nagu eelistas Bloomfield, või ideed ja mõisted inimese ajus, nagu arvas Locke? Ja kuidas seda tähendust seletada? Leksikaalsed teadmusbaasid keskenduvad tähenduse kirjeldamisele, võttes esimesena ette arvutisõnastikes olevad sõnade tähenduste seletused. Keele võlu ja "õnnetus" on, et (üht) keeleüksust seletatakse (teise) keeleüksuse abil – nii on see ka sõnaraamatus. Selge on see, et seletustega arvutisõnastik annab keele mõistmisse oma osa, aga see on vaid osa, ning ükskõik kui võimsaks ja peeneks arenevad sõnastike analüüsimise mehhanismid, tegelikkus ja inimese aju tegevus jäävad alati sõnastikust väljapoole.

2. Sõnaraamatu info analüüs

R. Amsler on terminit arvutileksikograafia (ka: arvutileksikoloogia) muuhulgas lahti mõtestanud kui arvutisõnastike igakülgset uurimist. On üldteada tõde, et sõnastiku tavakasutaja jaoks jääb suur hulk sõnaraamatuinfot varjatuks, et peidus on alati rohkem kui välja paistab. Arvutisõnastike puhul on ülitähtis töödelda implitsiitne info

[1] Kirevaid sõnastikufakte pakutakse leheküljel www.oed.com/inside/funfacts.html.

eksplitsiitseks, muundada nähtamatu nähtavaks. Selleks analüüsitakse sõnastikku ülima põhjalikkusega, tuvastatakse kõik võimalikud infoüksused ning püütakse luua mingid formalismid leitud üksuste kirjeldamiseks. Niisugune sõnaartikli loogilise struktuuri läbitöötamine ja formaliseerimine lisab uusi ligipääsuteid andmete juurde. Kui seni ollakse harjutud otsima vaid märksõna, siis nüüd võib keele sõnavara sõeluda mitme kandi pealt, lähtudes näiteks tuletusinfost, terminoloogilistest märgenditest, kollokatsioonidest, paljudest leksikaalsetest suhetest (hüponüümia, sünonüümia jm) või semantilistest väljadest (nõnda on itaalia keele andmebaasi valikuid iseloomustanud Calzolari 1995).

Sageli – eriti juhul, kui tegemist on traditsioonilise sõnaraamatu arvutisse sisestatud versiooniga – on materjali esituskuju selline, mida ükski arvutiprogramm kohe töödelda ei saa, materjali "söödavaks" tegemine on aga keerukas ja aeganõudev ülesanne. Et leksikaalne andmebaas rahuldaks paljude erinevate lingvistiliste teooriate infovajadusi, peaks selle sisemine korraldus olema niimoodi ehitatud, et lingvistiline info oleks teoreetiliselt neutraalne (D. Walker: polüteoreetiline) ja et seda oleks võimalik jagada mitmete erinevate keelerakenduste vahel.

2.1. Sõnaraamat kui struktureeritud tekst

Igaüks, kes sõnaraamatu avab, näeb, et selle sisu ei ole päris vabalt kulgev tekst, vaid et tekst on läbinisti korraldatud teatavatesse struktuuridesse. Oma olemuselt on sõnaraamat vastandlik: ühelt poolt sisaldab see hulgaliselt vabas vormis määramata pikkusega teksti, mis teeb info paigutamise tavapärasesse andmebaasi ebamugavalt keerukaks; teiselt poolt on sõnastik ise vägagi struktureeritud – enamgi kui tavapärane andmebaas (Boguraev 1997) –, hargnedes arvukatesse hierarhilistesse kihistustesse. Moodne arvutisõnastik on hierarhilise ehitusega struktureeritud tekstiandmebaas, mis ühendab endas nii andmebaasisüsteemide kui ka vabas vormis teksti tunnusjooni.

Sõnaraamatu struktuuri all mõistetaksegi harilikult andmete organiseerimist sõnastikus, aga keeletöötluses ka andmetele ligipääsu. Sõnaraamatu sees vaadeldakse harilikult kaht infostruktuuri: makro- ja mikrostruktuuri.

2.1.1. Makrostruktuur

Makrostruktuur ehk jämedam kord on **sõnaraamatu** korraldus ja sellega seotud probleemid, nagu sõnaloendi koostamine (sõnavalik), sõnade järjestamine, lemmade esitamine jm.

Sõnaraamatu makrostruktuur võib esiotsa jätta mulje, et sõnad "on reas" ja probleeme polegi. Ent isegi siin on asi tegelikult keerulisem ja lahendamist ootavad mitmed küsimused:

- kas iga märksõna peaks saama omaette sõnaartikli või peaks neid kuidagimoodi ühte lõiku kokku grupeerima?
- kas kõik märksõnad peavad olema ranges tähestikulises järjekorras või tuleks kasutada mitmeastmelist järjestamist nt tuletiste pesade jaoks)? kas alfabeetilist järjestust teha tähe või sõna kaupa?
- kui märksõnad on grupeeritud pesadesse, kas siis tuleks nad igal juhul välja kirjutada või võiks märksõnade korduvaid komponente asendada mingi sümbol? kui asendada, siis millise sümboliga ja kuidas täpselt?

"Eesti kirjakeele sõnaraamat" (EKSS) paneb näiteks ruumi kokkuhoiu mõttes ühte lõiku korduva esiosaga liitsõnad. Samas ei teki sellist pesa, kus koos oleksid *kõik* ühesuguse algusega liitsõnad, sest range alfabeetilise järjestuse tõttu katkeb liitsõnaloend ning vahele tulevad muud sõnad või teistsuguse liitepiiriga liitsõnad:

saama|-aasta ... -aeg ... -ahne ... -himu ... -himuline ... -koht ...
saamaline ...
saama|mees ... -päev ...
saamatu ...
saamatult ...
saamatus ...
saamatuse|tunne ...
saamatuvõitu ...
saama|tüli ...
(EKSS)

Eksplitsiitsuse nimel tuleks kõik lekseemid täielikult välja kirjutada ja esitada üksnes täiskujul: kõik otsingumootorid töötavad libedamalt, kui on olemas koht, kus märksõna (või ühend) on kirjas ilma igasuguste leksikograafiliste ja grammatiliste lisamärkideta, st nii, nagu see algvorm tavatekstis esineb:

kodu|tuvi ... --uurija	(EKSS: so. kodu-uurija)
jää_kirme[tis]	(ÕS99: so. jääkirme, jääkirmetis)
lak(k)ekrants	(FRAS: so. lakekrants, lakkekrants)
kasv|aa ... k:aminen ... K.	(Perussanakirja: so. kasvaminen)

Sellised paberkujulise sõnaraamatu juurde kuuluvad lühendamiskombed peaksid jäämagi ainult paberile: vajalikud lõiked teeb keelesüsteem.[2] Sõnastike märgendamisel soovitataksegi kasutada kaht (või mitut) erinevat märgendit, üht "puhta" märksõna, teist lisamärkidega märksõna jaoks.

2.1.2. Mikrostruktuur

Mikrostruktuur ehk peenem kord on **sõnaartikli** korraldus ja puudutab kogu infoesitust sõnaartikli sees.

Sõnaraamatu mikrostruktuuri uurimisel eristatakse kõik vajalikud infotüübid ja andmeväljad ja esitatakse sõnaraamatu/sõnaartikli sisumudel – kõigi infotüüpide süsteemne kirjeldus. Sõnaartikli formaliseerimiseks kasutatakse tavaliselt mingit andmemudelit (tekstimudelitest tuleb lähemalt juttu allpool). Arusaamatusi võivad põhjustada sõnaartiklite individuaalsed ja tihti iseäralikud koostamisformalismid, samuti ollakse sageli erineval teoreetilisel seisukohal selle suhtes, millist infot (süntaktilist, semantilist ja pragmaatilist) üldse lugeda lingvistiliselt oluliseks ehk relevantseks, mille õige koht on formaliseeritud arvutisõnaraamatus. Tihti on erinevad sõnastikud tehtud erimoodi isegi ühe maja piires – sõltudes ajast, vahenditest, oskustest või muust –, nii võtab oma aja ka süvenemine erinevatesse sõnastikeformaatidesse.

Sõnaartiklis leiduva info kirjeldamiseks võib kasutada mitut erinevat abstraktsioonitasandit. Informaatiline lähenemine uurib eraldi info kolme komponenti – vormi, funktsiooni ja tähendust; keeleteaduslik lähenemine analüüsib sõnaraamatut lingvistiliste infotüüpide kaupa.

---

[2] Seesugune otsus ongi tehtud näiteks soome Perussanakirja täienduste tegemisel, kus alates 1998. aastast kasutatakse SGML-toega süsteemi FrameMaker ja iga märksõna saab endale omaette sõnaartikli ja on alati välja kirjutatud. Kuidas see kõik paigutada raamatusse, see pole hetkel oluline, tegeldakse andmebaasikirjete ja seejärel CD-ROM-ina ilmuvate täiendustega.

2.2. Vorm, funktsioon, tähendus

Nende üldiste kategooriate abil iseloomustatakse paljusid erisuguseid objekte, nende omadusi ja mõtet. Niisamuti nagu keelemärgi ehk tähendusliku keeleüksuse, nii ka terve sõnaraamatu puhul vaadeldakse eraldi infoüksuse formaalset väljendust (vorm), kombineerumist teiste elementidega (funktsioon) ja konkreetset sisu (tähendus).

Vorm on keeleüksuse väline kest – sõna kirjakuju (kirjapilt, Viks 1992: kirjutusviis) või hääldusku ju –, mis näitab, kuidas sõna kirjutatakse või häälda takse. Kirjakuju koosneb tähe-, hääldusku ju foneemijärjendist. Sõnaraamatu puhul tegeldakse enamjaolt marksõna vormiinfoga: õigekiri, liitsõnapiir, silbitus või poolitus, hääldus, välde, rõhk jm. Antakse rohkem või vähem rikkalikult infot nii märksõna kui ka selle variantide kohta:

etalon e. **etaloon**	(EKSS)
('ões) 'õekse[d	(VVS: sulgudes on mittekasutatav variant)

Funktsioon kirjeldab keeleüksuse distributiivset käitumist, esinemistingimusi teiste keeleelementide suhtes. Sõnaartiklis võidakse piirduda ainult sõnaliigi märkimisega, kuid võidakse esitada ka peen grammatiliste kategooriate süsteem:

himur ... adj.	(EKSS)
threaten Ww4,5;T3	(LDOCE: Ww4,5 on verb, mis võib esineda ka adjektiivselt (*-ed, -ing*); T3 on transitiivne verb, millele järgneb *to*-infinitiiv.)

Ilma grammatilise infota ei saa läbi ükski keeletöötlussüsteem, aga just see info võib üldsõnastikes olla hõredamini esitatud, õigemini: vähem eksplitsiitselt välja toodud. Oleneb muidugi keelest – "ilma morfoloogiata" inglise keele puhul saab palju suuremas ulatuses toetuda üksnes tavasõnastiku infole kui näiteks eesti keele suguse keeruka morfoloogiaga keele puhul.

Tähendus kirjeldab keeleüksuse mõtet või sisu. Sõnaraamatus toimub see harilikult tähenduste seletuste ja näidete varal. Leksikaalse tähenduse – mida kannab sõnatüvi – kõrval on hulgaliselt juhtumeid, kus üksikul muutevormil on tüvest pisut erinev, omaette leksikaalne tähendus ja teda kirjeldatakse sõnaraamatus kui pooleldi iseseisvat, hrl alamärksõna:

kuulma ... 5. kõnek. **kuule** v. **kuulge** teat. kõnetlusvormel. (EKSS)

Mõned sõnaraamatud pakuvad sõna kasutuse täpsustamiseks rohkelt lisainfot, nii grammatilist kui ka stilistilist, sünonüüme ja antonüüme, erinevaid viiteid. Tähenduse paremaks selgitamiseks võidakse kasutada ka pildimaterjali, etümoloogilist infot jm. Tihedalt tähendusekomponendiga seotud on tuletised, liitsõnad ja fraasid (ühend- ja väljendverbid, fraseologismid, kollokatsioonid), mis otseselt sõna tähendusse ei kuulu. Sagedasti on need esitatud eraldi plokkidena sõnaartikli lõpus (alamärksõna, allkirje, *run-ons*).

Piirjooned kolme komponendi vahel ei ole aga sugugi nii selged, nagu eelnevast paista võib. Sõna tähendust võivad määrata grammatilised iseärasused, tähendusega on väga lähedalt seotud, ometi otseselt tähendusse kuulumata, sõna kombinatoorika – liitsõnadest kuni püsiühenditeni välja.

2.3. Lingvistilised infotüübid

Lähtuvalt keelekirjelduste tasanditest jagatakse sõnaartiklis leiduv info 6 laiaks infotüübiks:

- ortograafiline ehk grafoloogiline info (hrl märksõna),
- foneetiline info[3] (hrl hääldus),
- grammatiline info (morfoloogiline, sõnamoodustus- ja süntaktiline info),
- pragmaatiline info (hrl stiili-, eriala-, territoriaalsete või sotsiaalsete murrete märgendid; märkused),
- semantiline info (hrl tähenduste seletused, sünonüümid, tuletised, etümoloogia),
- illustreeriv materjal, mille eesmärk on näidata sõna kontekstis.

Enamik indoeuroopa sõnaraamatuid pakub vähemalt 3 tüüpi infot: foneetilist, süntaktilist ja semantilist (Boguraev 1997). Eestlastel on häälikkiri, mistõttu eesti sõnastikes ei ole foneetiline info sugugi nii

[3] Foneetika ehk häälikuid uuriv teadus ei taha hästi keeleteaduse alla mahtuda – sel on rohkem kokkupuutepunkte anatoomia, füüsika ja psühholoogiaga. Keeleteadus algab alates fonoloogiast ehk foneemi tasandist. Kuivõrd sõnaraamat on keeleasi, niivõrd jäävad keeleasjaks ka sõnaraamatutesse kirja pandud hääldamisjuhised. Ingliskeelses kirjanduses nimetatakse häälduslnfot sageli fonoloogiliseks (*phonological*), st keeleteaduslikuks infoks.

tähtsal kohal – hääldus võib peaaegu kõrvale jääda nii ükskeelses sõnastikus (EKSS) kui ka kakskeelse sõnastiku eesti pooles (eesti–vene, eesti–inglise), piirdudes vaid vähemate tsitaatsõnadega. Samas ei pea eesti ükskeelsedki sõnastikud paljuks märkida kaashäälikute peenendust ehk palatalisatsiooni (ÕS 76, ÕS 99), murdesõnastikust rääkimata (EMS). ÕS 76, VVS ja ÕS 99 märgivad nii väldet kui (ebaregulaarset) sõnarõhku, kuid need on juba otsapidi morfofonoloogiasse kuuluvad nähtused (välde eristab sõnavorme). Võib vist öelda, et eesti sõnastikes hõivab foneetilise info koha teine infotüüp – morfoloogiline info –, mis suunab sõna õigesti muutma või üldse mitte muutma. (Lingvistiliste infotüüpide lähem vaatlus jääb selle artikli piirest välja.)

3. Sõnaraamatu info kirjeldamine

Kuidas kirjeldada keelt nii, et arvuti sellest aru saaks? Kirjeldamise abivahend peab objektkeele väljendeid (märke, lauseid, fraase, sõnu jm) **nimetama**, peale selle aga ka iseloomustama nende väljendite **süntaksit** – moodustamise ja teisendamise reegleid – ja **semantikat** – suhet sellega, mida nad tähistavad. Tegevust, mille käigus valmib sõnastiku täielik struktuurikirjeldus (andmemudel) ja töötatakse välja spetsiaalne kirjelduskeel, nimetatakse sõnaraamatu formaliseerimiseks. Selle tegevuse eesmärk on teha sõnaraamatu tekst automaatselt töödeldavaks.

Sõnastikes leiduva tohutu info korraldus ei puuduta mitte ainult andmete säilitamist, vaid ka nendele hõlpsat juurdepääsu. Algteksti teisendamine mingisse metaformaati (andmebaas, andmemudel) on parasjagu keeruline ülesanne. Ei ole ka üheselt selge, milline metaformaat on parim (Boguraev 1997). Oma sõna on öelda ka tarbijal – andmebaas on kasutu, kui see ei suuda teenida ei inimkasutaja ega keeletöötluse huve. Arvutileksikograafia igavene "nuhtlus" on, et ühelt sõnastikult eeldatakse kaht erisuunalist asja: trükis ilmutamist ja andmebaasiks olemist.

Kui põhjaliku infoga leksikaalne üksus varustada – see sõltub ennekõike endale võetud ülesandest. Hõlpsamini formaliseeritavad on õigekiri, fonoloogia ja morfoloogia, ka sõnade süntaktiline kirjeldus (st vormi- ja funktsiooniinfo), määratult raskem on ühtseid lahendusi leida süvakäänete, semantika ja pragmaatika kirjeldamiseks, tegelikkuse hõlmamisest rääkimata. Realistliku ja representatiivse leksikoni koostamine on väga raske ülesanne, sest puudub

piisavalt põhjendatud teooria, mida see peaks sisaldama ja käsitlemist nõudvate sõnade hulk on tohutu suur. Leksikoni formaliseerimisel tuleb silmas pidada ka seda, et keeleüksuse kõiki tunnuseid ei ole võimalik või vajalik eksplitsiitselt esitada – paljud tunnused laienevad automaatselt ülemklassilt alamklassile (ingl *inheritance*, ee pärilus), nii nagu sõna *mees* on substantiiv ja muutub ühtmoodi nii 1., 2. kui 3. tähenduses.

Ideaalne sõnaraamatu "ehitaja" (*lexicon builder*) peaks niisiis olema arvutilingvist, formaallingvist, leksikograaf, tõlkija ja korpuselingvist ühes isikus. Samuti ei teeks paha tunda veidi ka infoteadust. Ideaali poole saab ainult püüelda. Reaalses elus on teinekord targem teisi mooduseid leida, näiteks komplekteerida arvutileksikograafiline töögrupp eri ampluaaga inimestest. Viimasel ajal on arutlusainet andnud ka tõsiasi, et elektrooniliste sõnastikega tegelevad inimesed, kel leksikograafiline ettevalmistus puudub täiesti (Kilgarriff 1999). Võivad tekkida vastuolud sellel pinnal, et arvutiinimestel ja leksikograafidel on sõnaraamatust erinev pilt, esimeste jaoks on see suvaliste sümbolite jada, teiste jaoks viimse kui detailini läbi mõeldud tekst, milles ükski kirjavahemärk, kirjastiil, tühik ega lõiguvahe pole ilma tähenduseta.

Sõnaartikli ehk sõnastiku mikrostruktuuri süsteemsel kirjeldamisel tuleb otsustada:

- millises formaadis esitada sõnaartikli eri osad ehk struktuuriüksused,
- kui eksplitsiitseks peab saama sisemine esitus,
- kas tahetakse kujundada võimsat andmebaasi või saadakse hakkama teksti märgendamisega, mis koos suhteliselt hõlpsate programmidega on nö andmebaasi simulatsioon.

Arvutilingvistiline uurimistöö ei järgi alati lingvistilisi kaanoneid ja loob arvutimudeleid, mis otseselt ühegi lingvistilise teooriaga seotud ei ole, kuid mis aitavad töödelda leksikoni sisu ja struktuuri (Boguraev 1997). Tuntumad andmete kirjeldustüübid leksikograafias on tekstimudelid, kontekstivaba grammatika, relatsioonilised andmemudelid ja tunnustel põhinevad andmemudelid (Ide jt 1997).

3.1. Tekstimudelid

Tekstimudelid on andmekogud, mis koosnevad peaasjalikult ainult tekstist või tekstide kogust. Lisainfot, mida vajavad kõik teksti ja

sõna töötlevad süsteemid, antakse eriti varasemates tekstimudelites õige napilt. Füüsiliselt kujutab tekstimudel endast järjestikust teksti, mida katkestavad mitmesugused abisümbolid. Arvutiprogrammid peavad suutma seda teksti "lugeda", st eristada sümbolijadast vajalikku infot. Selleks vajavad programmid aina põhjalikumat lisainfot teksti struktuuri ja sisu kohta ning üheks enamlevinud lisainfo andmise viisiks ongi sõnastikuteksti märgendamine (*markup, tagging*), mille areng on kulgenud lihtsatest kirjastiili käskudest teksti loogiliste süvakihtide tähistamiseni.

3.1.1. Tüpograafiline märgendus

Tüpograafiliselt märgendatud teksti sees olevad lisasümbolid kirjeldavad ainult teksti tüpograafilist külge: poolpaksu kirja, kursiivi jt kirjastiile ning teksti asetust. See on nn **toiminguma**rgendus (*procedural*), mis märgendite abil reguleerib kirjastiilide toimimist tekstis (tähistatud on käsu kehtimise algus ja lõpp). Mõnikord on seda nimetatud ka **kujundlikuks** (*presentational*) märgendamiseks. Teataval määral avab toimingumärgend ka pealispinna all olevat sisu – mis on poolpaks, on märksõna jne –, aga ainult teataval määral. Tavaliselt on eristatavaid infoüksusi kaugelt enam kui jätkub harjumuspäraseid kirjastiile: poolpaksud on EKSS-is lisaks märksõnale ka tähendusnumber, homonüümiainfo, ühendverbid ja -sidesõnad, fraseologismid, tähendusnihkega sõnavormid.

Tüpograafilisi käske ei tohiks õigupoolest märgenduseks lugedagi, sest see ei tegele üldse loogiliste tekstielementidega ning ei võimalda teksti muud töötlemist. Teksti komponentide täielik algoritmiline eristamine on võimatu. Infootsing niisuguses tekstis nõuab ilmselgelt palju pingutusi. Kuna aga enamik sõnastikesüsteeme on lähtunud trükimärkidega magnetlindist või tekstifailist, siis on neid käske n-ö jõuga käsitatud kui algelist märgendust:

> <B>5.<D><P8> hrv.<DP255> nägemine.<I> Küll öösel unes nägin / üht rasket nägu ma: / üks lillep<177><177>sas kasvas / mu aias üksinda.<D> L. Koidula.<B> <D>||<B><N><D><B>näoks, näo pärast<D> teistele nägemiseks; moepärast.<I> Tööd tehti ainult näo pärast. Metsa k<177>ndima minnes v<177>tsin näoks marjakorvi kaasa. <D><195><MI><N>Ja egas mina kuigi palju <D>[ei sööl<I>. Ma niisama näo pärast, et k<177>htu petta.<D> F.<N>Tuglas.<B> <D>||<B><N> <D><B>nägudeni<D><P8> k<177>nek.<DP255> nägemiseni.<I> Nägudeni (siis)! Jääme nägudeni.
>
> (EKSS: *nägu* 5. täh., kus <B> on poolpaks, <D> on käsu lõpp, <P8> on 8-punktine kiri, <I> on kursiiv, <177> on õ, <N> on mitte-ridavahetav tühik, <195> on ümartäpp.)

3.1.2. Deskriptiivne märgendus

Polügraafiliste märkidega võrreldes on deskriptiivne (*descriptive*) märgendus liikunud sügavuti ja kirjeldab vähemalt osaliselt teksti sisu: tähistatakse kõik struktuuriüksused, ära märgitakse üksuste algus. Iga uus üksus lõpetab automaatselt eelneva. Teksti lisatud sümbolid on nagu sildiks või pealkirjaks järgnevale sisule – kirjeldavad struktuuriüksusi nende sisust lähtuvalt. Selle asemel et öelda "7-punktine kiri" öeldakse "sõnaliik" või "erialamärgend" või "grammatiline kategooria" Algsemalt tüpograafiliselt kujult deskriptiivsele üleminekuks on tulnud välja töötada spetsiaalsed teisendusreeglid ja -programmid.

Alates 4. vihikust (ilmus 1991) kasutatakse ka EKSS-i sisestamiseks arvutisse 1980. aastate keskel välja töötatud deskriptiivset märgendust. Sõnastiku koostamiseks ja sisestuseks kasutatakse tavalist tekstiredaktorit:

> i+5. u+hrv. t+nägemine. n+Küll ööse1 unes nägin / üht rasket nägu ma: / üks lillepõõsas kasvas / mu aias üksinda. o+L. Koidula. i+|| m+näoks, näo pärast t+teistele nägemiseks; moepärast. n+Tööd tehti ainult näo pärast. Metsa kõndima minnes võtsin näoks marjakorvi kaasa. *Ja egas mina kuigi palju [ei sööj. Ma niisama näo pärast, et kõhtu petta. o+F Tuglas. i+|| m+nägudeni u+kõnek. t+nägemiseni. n+Nägudeni (siis)! Jääme nägudeni.
>
> (EKSS: *nägu* 5. täh., kus i+ on tähendusindeks, u+ on stiilimärgend, n+ on näide, o+ on tsitaadi autor, m+ on märksõna, t+ on tähendus, * eraldab leksikograafi näitelauseid tsitaatidest.)

Samal moel on EKI-s tehtud ka mitu teist sõnaraamatut, värskeim näide on 1999. aastal ilmunud ÕS 99. Deskriptiivset märgendust kasutab ka loodav etümoloogilise kartoteegi arvutiarhiiv ning samal põhimõttel sisestatakse ka Wiedemanni sõnaraamatut:

> m+laenama
> g+-nan, -nata (nada) 1v
> v+laenatama, lainama
> t+leihen, borgen,
> y+ära l\aenama\, väl'ja l\aenama\
> t+ausleihen,
> f+hinge ära laenama
> t+sterben.%
>
> (WIED: *laenama*, kus y+ tähistab ühendit ja längkriipsude vahel on raamatu lühenduse eksplitsiitne väljakirjutus, muude märkide tähendust vt ülalpool.)

Deskriptiivse märgenduse puuduseks on tema lamedus ehk lineaarsus, kajastamata jääb sõnaraamatu kui struktureeritud teksti

olemus – sõnaartikli hierarhia. Ülaltoodud EKSS-i näiteski hargneb sõna *nägu* 5. tähendus kaheks omaette tähendusvarjundiks, millel kummalgi on oma vorm (*näoks, näo pärast* ja *nägudeni*), oma tähendus ja ühel ka märgend (*kõnek.*). Sõnaraamatus on keerulisi hierarhilisi struktuure, kus vaja võib minna 5–6 erinevat kihistust. Deskriptiivne märgendus seda ei võimalda.

3.1.3. Üldistatud märgendus

Üldistatud (*generalized*) märgendamise põhimõtted postuleeriti juba 1970. aastatel: a) märgendamine peab kirjeldama teksti struktuuri, mitte hilisemat teksti töötlemist; b) märgendatud tekst peab olema algoritmiliselt töödeldav. Üldistatud märgenduskeele rahvusvaheliseks standardiks kinnitati 1986. aastal *Standard Generalized Markup Language*, lühendatult SGML (ISO Standard 8879). SGML on uudne infokeel, mis ei ole seotud ühegi konkreetse tarkvaraga, st on vaba nende poolt dikteeritud tingimustest ning seega põhimõtteliselt avatud paljudele erinevatele süsteemidele. Ka peaks SGML ajale paremini vastu pidama kui masinate arengust sõltuv ja kiirelt uuenev tarkvara.

Üldistatud märgendus nõuab nii struktuuriüksuse alguse (*<tähendus>*) kui ka lõpu (*</tähendus>*) tähistamist, misläbi muutuvad eksplitsiitseks üksuste omavahelised alluvussuhted: märgenduse toel saab ära kirjeldada kogu teksti loogilise ülesehituse. Graafiliselt kujutatakse märgendatud tekstide hierarhilist ehitust puukujulise struktuurina. Kui SGML-i võrrelda kahe eelmainitud märgendusviisiga, siis on see pigem analüütiline ja deskriptiivne, mitte konkreetsest väliskujust lähtuv või konkreetsele toimingule suunav. Üldistatud märgendus on ühtlasi esimene abivahend arvutisõnastike maailmas, mille abil lahendatakse ühe korraga kaks erimõõtmelist ülesannet: see võimaldab nii praktilist kui ka akadeemilist väljundit, st arvutisõnastikku saab nii trükis ilmutada kui ka leksikaalse tekstiandmebaasina kasutada.

Andmestruktuuride formaalne esitus pannakse kirja (unikaalse) SGMLi "dokumendi tüübideklaratsiooni", lühendatult DTD (*document type declaration*) abil, mis:

- lähtub teksti sisust (mitte lehekülgedest);
- kirjeldab teksti loogilist ehitust (mitte füüsilist).
- võimaldab märgendeid grupeerida arusaadavaisse gruppidesse;
- on avatud täiendustele ja isegi uute märgendite lisamisele.

Mõiste **dokument** ei ole füüsiline objekt (fail, peatükk), vaid loogiline konstruktsioon, mis hõlmab nii sisulisi elemente kui ka märgendust. Märgendus on **lisainfo** tegelike andmete juures, mille abil eristatakse erinevaid elemente. Üldistatud märgenduse abil muutub dokument kindla grammatikaga loogiliseks avaldiseks (Goldfarb 1990). DTD defineerib dokumendi elemendid ja elementide atribuudid koos nende võimalike väärtustega.

Element on dokumendi struktuuriüksus, millel on oma identifikaator (siin: *raamat, entry*) ja mille sisu avab loogilisi operaatoreid sisaldav regulaarne avaldis:

<! ELEMENT	raamat	(esiosa, põhitekst, järelosa) >, st raamat on üksus, mis koosneb fikseeritud järjekorras esiosast, põhitekstist ja järelosast.
<! ELEMENT	entry	(hom | sense | def | eg | form | gramGrp | note | re | trans | xr)+) >, st entry 'kirje' on üksus, mis võib sisaldada üht või mitut homonüümi, tähendust, seletust, näidet, märksõn, grammatilise info üksust, märkust, allkirjet või viidet.

Elemendi atribuut täpsustab või kitsendab elementi, iseloomustades selle teatud tunnusjooni. Atribuutide abil hoidutakse ka patustamisest Occami habemenoa vastu. Nt märksõna atribuudid võivad olla:

<form type=foreign> ehk tsitaatmärksõna
<form type=variant> ehk märksõna variant
<form type=compound> ehk liitsõnast märksõna
<form type=derivative> ehk märksõna tuletis
<form type=phrase> ehk mitmesõnaline leksikaalne üksus märksõnana

Kogu dokumendi kirjelduse õigsust kontrollitakse spetsiaalse SGML-parseri abil.

Üldistatud märgendus ehk SGML on arvutileksikograafias laialt pruugitud, seda kasutavad näiteks *Oxford English Dictionary*, Longmani, Collinsi jpt sõnastikud. Paljud kasutavad oma tarbeks tehtud mugandusi:

<Entry> <Head><HWD>zealot</HWD>
<HYPHENATION>zea.lot</HYPHENATION>
<PronCodes><PRON>'zel@t</PRON></PronCodes>
<POS>n</POS>
<GRAM>C</GRAM></Head>
<Sense><DEF>someone who has extremely strong beliefs, ...</DEF>
<EXAMPLE>religious zealots</EXAMPLE></Sense>
<Tail><RunOn><SPELLING>zeolotry</SPELLING>
<POS>n</POS>

```
<GRAM>U</GRAM></RunOn></Tail><Entry>
<Entry> <Head><HWD>bread</HWD>
<FREQ>S2</FREQ>
<FREQ>W3</FREQ>
<PronCodes><PRON>bred</PRON></PronCodes>
<POS>n</POS>
<GRAM>U</GRAM></Head>
<Sense><DEF>a common important food made from flour, water and
<NonDV><REFHWD>yeast</REFHWD></NonDV></DEF>
<EXAMPLE>Would you like some bread with your soup?</EXAMPLE>
<ColloExa><COLLO>a loaf of bread</COLLO>
<GLOSS>a large piece of bread that you buy and cut into
pieces</GLOSS></ColloExa>
<ColloExa><COLLO>a slice of bread</COLLO> ... </ColloExa>
<ColloExa><COLLO>white/brown bread</COLLO> ... </ColloExa>
</Sense>
<Sense><LITTLEWORDS>old-fashioned</LITTLEWORDS>
<DEF>money</DEF></Sense>
<Sense><LEXUNIT>your/sb's bread and butter</LEXUNIT>
<LITTLEWORDS>informal</LITTLEWORDS>
<DEF>...</DEF>
<EXAMPLE>...</EXAMPLE></Sense>
<Sense><LEXUNIT>...</Sense>
<Tail><Crossref><CROSSREFTYPE>--see also</CROSSREFTYPE>
<Crossref2><REFHWD>french bread</REFHWD></Crossref2>
<Crossref2><REFHWD>sliced
bread</REFHWD></Crossref2></Crossref></Tail></Entry>
```

(LDOCE (näited saadud A. Kilgarriffilt 1998. a): *Entry* – kirje, *Head* – (kirje) päis, HWD – märksõna, *Hyphenation* – poolitus, PRON – hääldus, POS – sõnaliik, GRAM – grammatilised andmed, *Sense* – tähendus, DEF – tähenduse kirjeldus või definitsioon, EXAMPLE – näiteväljend, *Tail* – (kirje) järelosa ('saba'), *RunOn* – tuletis, *Spelling* – (tuletise) kirjapilt, FREQ – sagedusandmed, REFHWD – viidatav märksõna, COLLO – kollokatsioon, GLOSS – (kollokatsiooni) tähenduse kirjeldus, LITTLEWORDS – väikses kirjas stiili- jm märgendid, LEXUNIT – mitmesõnaline leksikaalne üksus, *Crossref* – viide.)

Collinsi kirjastus demonstreerib avalikult oma kakskeelsete sõnastike märgendamist prantsuse–inglise sõnastiku näitel. Ka siin toetub märgendus SGMLile, ent seda on trükkimiseks veidi täiendatud:

```
<COMMON>

<HWME> eau
<HWAD> x
<PRON> o
<POSP> nf
<TRAN> acqua

<BFORMAT>

<PHRS> sans ~
```

```
<LBIN> whisky etc
<TRAN> liscio*
```

(Collinsi sõnastik: <COMMON> – grupeeriv märgend (esineb ainult andmebaasis), <HWME> – põhimärksõna (poolpaks kiri), <HWAD> – muutelõpp (poolpaks kiri), <PRON> – hääldus (nurksulgudes, harilik kiri, märksõna järel), <POSP> – sõnaliik (kursiiv), <TRAN> – tõlkevaste (harilik kiri), <BFORMAT> – grupeeriv märgend (esineb ainult andmebaasis), <PHRS> – ühend (poolpaks teises šriftis kiri), <LBIN> – tähendusvihje (kursiiv, ümarsulgudes).)

Seni ainsaks eesti näiteks saab tuua soomlastega kahasse tehtava soome–eesti sõnaraamatu. SGML-märgenduses algtekst saadi Helsingi *Kotimaisten kielten tutkimuskeskus*elt, ent kuna see oli soome–rootsi sõnastiku materjal, siis esimese ülesandena puhastati see rootsi keelest ja selle rootsikeelsed märgendid kohandati vastavalt meie tarbele. Eesti koostajad said oma töölauale (arvutisse) sõnaraamatu struktureeritud tooriku, mida seejärel sisuga täitma hakati:

```
<kirje>
<art>
<päis><ms>udmurtin kieli</ms></päis>
<täh>
<viide>vt udmurtti 2</viide>
</täh></art></kirje>
<kirje>
<art>
<päis><ms>ufo1</ms></päis>
<täh nr=1>
<tlk><info>&subst;</info>
<vst>...</vst>
</tlk>
</täh>
<täh nr=2>
<tlk><info>&slg; &adj;: outo</info>
<vst>...</vst>
<vst>...</vst>
</tlk>
<tlk><info>hullu</info>
<vst>...</vst>
</tlk><romb><nt>
<sm>tämähän on ufo juttu</sm>
<ee>...</ee>
</nt>
</täh></art></kirje>
```

(soome–eesti: <kirje> – sõnaartikkel, <art> – ühe märksõna kirjeldus, <päis> – sõnaartikli päis, <ms> – märksõna, <täh nr=...> – (1., 2. jne) tähendus, <tlk> – tõlkevaste koos oma infoga, <info> – märksõna grammatiline vm kitsendav info, <vst> – vaste, <nt> – näiteplokk, <sm> – soomekeelne näide, <ee> – eestikeelne näide, <viide> – viide.)

3.1.4. TEI

SGMLi üheks edasiarenduseks on rahvusvahelise uurimisprojekti *Text Encoding Initiative* (TEI) poolt välja töötatud valmis kodeerimisskeemid ehk märgendusmudelid (*tag sets*) paljude erinevate tekstitüüpide jaoks. Standardseid mudeleid pakutakse nii korpusetekstide kui ka trükitud sõnaraamatute märgendamiseks, mudeli päises (*header*) mainitakse täpselt, mis komplekti on kasutatud.

Sõnaraamatukirjet nagu süntaksitki kirjeldatakse hierarhilise moodustajastruktuuri abil. Hierarhiliselt kõige ülem struktuurielement sõnaartiklis on kirje (*<entry>*), mis järgmisel tasandil hõlmab homonüümide (*<hom>*), eri tähenduste (*<sense>*) ja (märksõnaga seotud) allkirjete (*<re>*) elemente.

```
<entry>
<!-- homonüümide kohta käiv ühine info: märksõna, grammatika jm -->
<hom n='1'>
  <sense n='1'>...</sense>
  <sense n='2'>...</sense>
</hom>
<hom n='2'>
  <sense n='1'>...</sense>
  <sense n='2'>...</sense>
</hom>
</entry>
<entry>
<!-- märksõna, grammatiline jm info (tähenduste ja allkirjete jaoks ühine) -->
  <sense n='1'>...</sense>
  <sense n='2'>...</sense>
  <sense n='3'>...</sense>
<re>
<!-- allkirje: fraseologism koos oma infoga -->
</re>
<re>
<!-- allkirje: ühendverb koos oma infoga -->
</re>
</entry>
```

Järgnevad kõrgema ja fraasitasandi moodustajad. Kõrgema tasandi moodustajad (*top-level constituents*) on: märksõna vormiinfo, grammatiline info, tähendusinfo, etümoloogiline info, näited, kasutusinfo, viited ja märkused. Need üksused võivad esineda igal ülemal struktuuritasandil (so. kirje, homonüümi, tähenduse ja allkirje tasandil), sest vastav info võib käia terve kirje kohta või ka üksnes ühe homonüümi või tähenduse või allkirje kohta. Kõrgema tasandi moodustajate erijooni – lähemaid omadusi ja olulisi tunnuseid –

esitatakse atribuutide abil, nii nagu SGMLgi seda teeb: näiteks "märksõnaks olev fraseologism" on *<form type=idiom>* või "mitmus" on *<gram type=number>pl.</gram>*. Fraasitasandi moodustajad (*phrase-level constituents*) on alamastme elemendid ("aatomid") – võrdväärsed vahetud moodustajad ülema struktuuri sees –, mis enam ise edasi ei hargne: näiteks "häältdus" (*<pron>*), "märksõna kirjakuju" (*<orth>*) jms.

TEI andmemudel elemendi *<entry>* ('kirje, sõnaartikkel') jaoks sobib tõenäoliselt paljudele erinevatele sõnastikele, ent ikka ja alati leidub sõnaartikleid või terveid sõnaraamatuid, mille struktuur selle mudeli sisse ei mahu. Näiteks võib seletuse sees ette tulla häältdus, mida range *<entry>*-mudel ei luba:

> **demigod** ... (in Gk myth, etc) the son of a god and a mortal woman, eg Hercules [häältdus]
> (Ide jt 1997)

Niisuguste juhtumite tarvis on TEI defineerinud vaba struktuuriga sõnaartikli mudeli *<entryFree>*, mis lubab suuri vabadusi ühe kirje sees: iga element võib esineda igal pool. Elemendi *<entryFree>* toel saab sõnastikku kirjeldada ilma ühegi grupeeriva elemendita, kasutades ainult fraasitasandi elemente. Tekib üleni lame sõnastikupilt, mis ei iseloomusta üldse hierarhilisi struktuure. On võimalik, et mingi ülesande lahendamisel see kodeerijat rahuldabki.

Ülaltoodud tekstijupp EKSSist näeb TEI märgenduses välja niimoodi (nagu näha, võtavad märgendid enda alla vähemalt sama palju ruumi, kui mitte rohkem, kui sõnaartikli tekst ise):

```
<sense n='5'>
<usg type=reg>hrv.</usg>
<def>nägemine.</def>
<eg><cit><quote>Küll öösel unes nägin / üht rasket nägu ma: / üks
lillepõõsas kasvas / mu aias üksinda.</quote>
<bibl>L. Koidula.</bibl></cit>
</eg>
<sense n='||'>
<form type="inflected">
<orth>näoks</orth>
<orth>näo pärast</orth>
</form>
<def>teistele nägemiseks; moepärast.</def>
<eg><quote>tööd tehti ainult näo pärast.</quote>
<quote>Metsa kõndima minnes võtsin näoks marjakorvi kaasa.</quote>
<cit><quote>Ja egas mina kuigi palju [ei söö]. Ma niisama näo pärast, et
kõhtu petta.</quote>
<bibl>F  Tuglas.</bibl></cit></eg>
</sense>
```

```
<sense n='||'>
<form type="inflected">
<orth>nägudeni</orth>
<usg type=reg>kõnek.</usg>
</form>
<def>nägemiseni.</def>
<eg><quote>Nägudeni (siis)!</quote>
<quote> Jääme nägudeni.</quote></eg>
</sense>
</sense>
(EKSS: nägu 5. täh)
```

EKSS on ka üks mitmest eri keelte sõnaraamatutest, millega mõõdetakse TEI märgendite universaalsust: 1998 alanud ühisprojekti CONCEDE raames üritatakse välja selgitada, mil määral alluvad ungari, sloveenia, tšehhi, bulgaaria, rumeenia ja eesti keele ükskeelsed sõnaraamatud ühele ja samale märgendusmudelile. Kui asi töötab, siis sünnivad aastaks 2001 kõigi nimetatud keelte sõnaraamatute ühtsed formaalsed andmemudelid.

3.2. Kontekstivaba grammatika

Kontekstivabade generatiivgrammatikate eeskujul on ka sõnaartikli lubatavate struktuuride kirjeldamiseks kasutatud kontekstivaba grammatikat. Kontekstivaba grammatika mudel koosneb reeglitest, sõnaraamatu hierarhilist olemust analüüsitakse nende reeglite ja puukujulise skeemi toel. Iga reegli vasakule poole jääb ainult üks elementaarüksus, näiteks *artikkel* → *märksõna näide* ("artikkel koosneb märksõnast ja näitest").

1980. aastate teisel poolel käivitus ka Eestis sõnastike andmebaaside ühisprojekt Keele ja Kirjanduse Instituudi (praegu EKI) ja Tallinna Polütehnilise Instituudi (praegu TTÜ) vahel, kus meetodiks oli samuti kontekstivaba grammatika. Kirjelduskeele ELMALEX abil kirjeldati EKSS-i struktuuri ehk süntaksit.

```
keel      -> leksikon start
leksikon  -> (artikkel %%)*
artikkel  -> (art)*
art       -> pea
          -> pea kesa
kesa      -> keha
          -> keha saba
          -> saba
pea       -> päis
          -> (päis gr,)#
päis      -> mtähis märks
          -> mtähis (märks mkvm)#
```

```
          -> mtähis (märks mkvm)*
mtähis    -> m+
          -> z+
          -> f+##
          -> f+
märks     -> mrk
          -> mrk (ka: mrk)
          -> mrk e. mrk
          -> mrk:
          -> fmrk
mkvm      -> .
          -> .
mrk       -> SÕNA
          -> COMP
          -> COMB
fmrk      -> mrk ~ (mrk ~)#
...
```

3.3. Relatsioonilised andmemudelid

Relatsiooniline andmebaas on organiseeritud andmekogum, mis fikseerib baasandmed ja nende suhted. Andmed on korraldatud väljadeks ja kirjeteks, millede vahel tekib seoste (ehk relatsioonide) tabel, kus igas lahtris on üks ja ainult üks väärtus. Enamasti kasutavad niisuguseid andmemudeleid arvutilingvistid, vähem populaarsed on need leksikograafide seas, kes kipuvad sellist mudelit pidama liiga lihtsustavaks ja ka liiga jäigaks. Filoloogiharidusega uurijale võivad need olla ka lihtsalt mõistetamatud. Siiski on mitmed uurijad on seda mudelit propageerinud ka leksikograafias kasutamiseks (Ide jt 1997).

Näiteks Collinsi kirjastuse leksikaalse andmebaasi aluseks on Illinois' tehnoloogiainstituudis välja töötatud Oracle'i relatsiooniline andmebaas. Andmed paigutatakse kihtidena suurtesse lameandmebaasi failidesse (*flat files*). Põhitabel esitab keeleüksuse ja selle identifitseeriva sümboli, kogu muu info paikneb kahes lisatabelis (Conlon jt 1997).

Eesti leksikograafias on see mudel samuti tuttav. Relatsioonilise andmebaasina on arvutis esimene eesti keele automaatsõnastik – "Õigekeelsussõnaraamatu" 1980. aastal valminud 114 000 märksõnaga arvutivariant. Kõik struktuuriüksused on organiseeritud positsioonide järgi (Viks 1981):

Märksõna	Tüübinr	Tüvevok	Sõnaliik	...	Märgend
,aadama+,aegne					P
aadeldama			V		T
,aadel/k,ond	55	A	S		
aadl[,ik	82	U	S		
aafrika			G		

(ÕS: on III välde, + on liitsõnapiir, / ja [on tuletusliite piirid.
Osa välju (uus tüübinumber ja tüvevokaal) on siinse ruumi kokkuhoiu mõttes esitamata, neid asendab kolm punkti (...).)

Ka 7500 sõna sisaldav slängisõnastik (Loog 1991) on arvutis relatsioonilise andmebaasi kujul:

Teema nr	Sõna	Päritolukeel	...	Tüdruk v poiss	Vanus
082	õgima	e		T	15
082	õgima	e		T	16
082	öökima	e		P	18
083	janu tegema	e		T	16
083	kuuli panema	e		P	17
083	lahendama	e		P	16

Ka EKSSi deskriptiivse märgendusega teksti põhjal on katsetatud programmeerida lameandmebaasi, vastavad programmid koostati TÜ arvutuslingvistika laboris. Sõnaartikkel lahutati 12 väljaks (millest mõned võivad täitamata jääda): tunnusnumber, märksõna (või mingi muu märksõna positsioonis olev üksus), hääldus, grammatika, sõnaliik, erialamärgend, stiilimärgend, tähendusindeks, seletus, näited, liitsõnad, viited. Tabelina näeks see välja nii nagu tabelis järgmisel leheküljel (ruumi kokkuhoiu mõttes on ära jäetud tunnusnumbri veerg, samuti selle konkreetse sõna puhul tühjaks jäänud veerud).

Mida keerukam on andmetekogu, seda keerukamaks lähevad tabelid. Et andmeid mitte liiga palju dubleerida, jaotatakse info mitmesse erinevasse tabelisse. Koondtabelid sisaldavad harilikult tohutult palju "üleliigset" infot (Ide jt 1997), aimduse saab siin toodud näitest. Suurimaks probleemiks ongi andmete kirevus ja mittelineaarsus: ühes sõnaartiklis võib olla mitu sõnaliigimärgendit, mitu seletust, mitu märksõna staatuses sõna.

Märksõna	Gramm	Sõnaliik	Stiil	Täh.ind	Seletus	Näited	Viited
m+viis,	g+viie, viit 35	s+num.		i+1.	t+põhiarv 5.	n+Viis pluss viis on kümme.	
m+viis,	g+viie, viit 35	s+num.		i+1. i+||	t+põhiarv 5. ==> t+(vastava arvulise järjekorra kohta).	n+Punkt, peatus viis t+'viis punkt, peatus'!	
m+viis,	g+viie, viit 35	s+num.		i+1. i+|	t+põhiarv 5. ==> t+(kellaja kohta).	n+Kell viis kakskümmend kaks minutit.	
m+viis,	g+viie, viit 35	s+num.		i+1. i+||	t+põhiarv 5. ==> t+hulgalt, koguselt 5.	n+Viis last, sõpra.	
m+viis,	g+viie, viit 35	s+num.	u+piltl.	i+1. i+||	t+põhiarv 5. ==> t+(käe kohta).	n+Viis pihku, oleme sõbrad!	
m+viis,	g+viie, viit 35	s+s.		i+2. i+a.	t+number 5.	n+Araabia, rooma viis.	
m+viis,	g+viie, viit 35	s+s.		i+2. i+b.	t+parim hinne viiepallises hindamissüsteemis.	n+Sai kirjandi eest viie.	
m+viis,	g+viie, viit 35	s+s.		i+2. i+c.	t+(muid juhte).	n+Poti viis t+'5 potisümboliga mängukaart'.	
f+läbi viie (sõrme) vaatama					t+millestki ilma jääma; millelegi tähelepanu mitte pöörama.	n+Maja põles maha, nüüd vaata läbi viie sõrme.	
f+nagu viis kopikat							e+vt. kopikas.

3.4. Tunnustel põhinevad mudelid

Kuna klassikalised andmemudelid (relatsiooniline andmebaas) ei ole väga hästi kohaldatavad keeruliste, tihti paljusõnaliste leksikaalsete andmete jaoks, siis on leksikaalse info kodeerimiseks propageeritud ka tunnuste struktuuride teoorial (*feature structures*) põhinevat andmemudelit. See mudel on mõnede uurijate sõnutsi lingvistilise info esituskujuna küllalt laialt levinud, samuti peaks see ühilduma kommertslike andmebaaside tarkvaraga (*database management system*, DBMS), ehkki seda polevat seni tehtud (Ide jt 1997). Mudel kirjeldab objekte ja nendevahelisi seoseid, st tegemist on objekt-orienteeritud mudeliga. Formaalses ja arvutilingvistikas on tunnustel põhinevad mudelid laialt tuntud, nende abil on kodeeritud lingvistilist infot, eriti grammatilisi formalisme. Sama loomulik tundub olevat kasutada neid sõnastike jaoks, kus esmapilgul täiesti ühesugusena tunduv info (ortograafia, hääldus, sõnaliik, etümoloogia, seletused jm) on esitatud mitmes eri kohas ja mitmel eri viisil.

Tunnuste mudeli aluseks on SGML-märgenduses sõnastikutekst. Iga struktuur (nt *form, gram, def*) vastab ühele SGML-elemendile (*<form>...</form>* jne).

Tunnuste struktuur koosneb tunnustest ja nende väärtustest. Kirjapildis eraldab tunnust väärtusest koolon, kokkukuuluvad liikmed on ühendatud nurksulgudega. Väärtuseks võib olla ka teine struktuur. Lihtne sõnaartikkel näeks välja nii:

s [ess], s-i[22] eesti tähestiku täht; vastav häälik. (EKSS)

```
[ form:   [ orth: s   ]
          [ pron: ess ]
  gram:     infl: s-i
            [ itype: 22 ]
  def:      [ text: eesti tähestiku täht; vastav häälik.] ]
```

Mitmetähenduslike sõnade esitus:

päri I. adv. **1.** (ettepaneku, väite vms. suhtes) samal arvamusel... *Otsusega päri olema.*
2. mingi liikumisega samas suunas; ant. vastu. *Tuul oli päri.*
3. van. soodsalt, kellelegi, millelegi vastavalt. *Olukord oli meile päri.*
II. prep. [part.] mingi liikumisega samas suunas; ant. vastu (sageli kirjutatakse järgneva sõnaga liitadverbina kokku). *Tüüris laeva päri tuult.*
III. postp. hrv. [gen.] **1.** mingi liikumisega samas suunas. Nüüd toob üks just tuule päri keeris kõrget tolmusammast piki maanteed ligemale. M. Raud.

2. millelegi sobivaks, vastavaks. *..ei ulatanud kätt mitte teretamise päri.* R. Põder.

```
form:          [ orth: päri ]
 // sense I
 gram: [pos: adv ]
  // sense 1
  def:         [ text: (ettepaneku, väite vms suhtes) samal arvamusel]
  ex:          [ text: Otsusega päri olema]
  // sense 2
  def:         [ text: mingi liikumisega samas suunas; ant. vastu.]
  ex:          [ text: Tuul oli päri.]
  // sense 3
  usage:[ reg: van.]
  def:         [ text: soodsalt, kellelegi, millelegi vastavalt.]
  ex:          [ text: Olukord oli meile päri.]
 // sense II
 gram:          pos: prep
                gov: part
 def:         [ text: mingi liikumisega samas suunas;
                ant. vastu (sageli kirjutatakse järgneva sõnaga
                liitadverbina kokku).]
 ex:          [ text: Tüüris laeva päri tuult.]
 // sense III
 gram:          pos: adv
                gov: gen
 usage:       [ reg: hrv.]
  // sense 1
   def:        [ text: mingi liikumisega samas suunas.]
   ex:         [ text: Nüüd toob üks just tuule päri keeris kõrget
                tolmusammast piki maanteed ligemale. M. Raud. ]
  // sense 2
   def:        [ text: millelegi sobivaks, vastavaks.]
   ex:         [ text: ..ei ulatanud kätt mitte teretamise päri. R. Põder.]
```

Tunnuste mudel on leksikaalsete andmebaaside jaoks iseenesest väga hea, sest lubab paindlikult kirjeldada suure ükskeelse sõnaraamatu kogu hierarhilist küllust. Väärtused on ühilduvad (*compatible*) või mitteühilduvad (*incompatible*) – erandeid käsitletakse näiteks kui mitteühilduvat infot: EKSS-i *nägu*-artiklis oleks erandiks tähenduse sees olevad märksõnavormid *näoks, näo pärast* ja *nägudeni*). Alltasandi tunnuse väärtus alistab (*override*) pealistasandi sama tunnuse väärtuse – juuresolevas näites alistab 3. tähenduse häälduse /hääldus2/ sõna tavapärase hääldus e /hääldus1/:

con.jure /hääldus1/ *vt, vi* **1** [VP2A,15A] do clever tricks... **2** ... **3** /hääldus2/ [VP17] (formal) appeal solemnly to...
(Ide jt 1997)

Tunnuste mudeli rakendamise teeb raskeks asjaolu, et pole olemas üldlevinud tunnuste mudelile toetuvat andmebaaside tarkvara.

Kirjandus

Amsler, R. 1980. The Structure of the Merriam-Webster Pocket Dictionary. Doctoral dissertation, University of Texas at Austin.

Boguraev, B. 1997. Machine-readable dictionaries and computational linguistic research. – Current Issues in Computational Linguistics: In Honour of Don Walker. Toim A. Zampolli, N. Calzolari, M. Palmer. Kluwer. 119–154.

Calzolari, N. 1995. Structure and access in an automated lexicon and related issues. – Automating the Lexicon: Research and Practice in a Multilingual Environment. Toim D. Walker, A. Zampolli, N. Calzolari. Oxford University Press. 337–356.

Conlon, S. P.-N., Dardaine J., D'Souza, A., Evens, M., Haynes, S., Kim, J.-S., Strutz, R. 1997. The IIT lexical database: dream and reality. – Current Issues in Computational Linguistics: In Honour of Don Walker. Toim A. Zampolli, N. Calzolari, M. Palmer. Kluwer. 201–225.

eesti–inglise = Eesti–inglise sõnaraamat. Ilmumas.

eesti–vene = Eesti–vene sõnaraamat I–II. Tallinn. 1997–2000.

EKSS = Eesti kirjakeele seletussõnaraamat. Tallinn. 1988–1999.

EMS = Eesti murrete sõnaraamat. I–II. Tallinn. 1994–1998.

FRAS = Õim, A. 1993. Fraseoloogiasõnaraamat. Tallinn.

Goldfarb, C. F. 1990. The SGML Handbook. Oxford University Press.

Ide, N., Le Maitre, J., Véronis, J. 1997 Outline of a model for lexical databases. – Current Issues in Computational Linguistics: In Honour of Don Walker. Toim A. Zampolli, N. Calzolari, M. Palmer. Kluwer. 283–320.

Kilgarriff, A., Rundell, M., 1999. Lexicography for computationalists. (Tutorial.) – Proceedings of EACL '99. Ninth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. xviii.

LDOCE = Procter, P (toim) 1978. Longman Dictionary of Contemporary English. Harlow and London: Longman.

Loog, M. 1991. Esimene eesti slängisõnaraamat. Tallinn.

Perussanakirja = Suomen kielen perussanakirja I–III. Helsinki. 1995.

soome–eesti = Soome–eesti sõnaraamat. Koostatav käsikiri Eesti Keele Instituudis.

Viks, Ü. 1981. Eesti keele automaatsõnastikest. – Keel ja Kirjandus 4, 211–216.

Viks, Ü. 1992. Väike vormisõnastik. Sissejuhatus ja grammatika. Tallinn.

VVS = Viks, Ü. 1992. Väike vormisõnastik. Tallinn.

Walker D., Zampolli A., Calzolari N. (toim) 1995. Automating the Lexicon: Research and Practice in a Multilingual Environment. Oxford University Press.

ÕS 76 = Kull, R., Raiet, E. (toim) 1976. Õigekeelsussõnaraamat. Tallinn: Valgus.

ÕS 99 = Erelt, T. (toim) 1999. Eesti keele sõnaraamat. Tallinn: EKSA.

Eesti keele tesaurus

Kadri Vider, Neeme Kahusk, Heili Orav, Haldur Õim, Leho Paldre
Tartu Ülikool

1. Sissejuhatus

Vajadus kaasaegse arvutitesauruse järele on eesti leksikoloogias ammune. Kui ilmus Andrus Saareste "Eesti keele mõisteline sõnaraamat" oli eestikeelne, ainult arvutis teostatud tesaurus veel mõeldamatu. Ka "Eesti kirjakeele seletussõnaraamatut" hakati koostama eelkõige paberil väljaandmiseks. 1980. aastate lõpul tehti Eesti Keele Instituudis esimesi katseid luua tesaurust automaatselt, olemasolevate sünonüümiridade põhjal (Hallik 1990). 1997. aastal loodi Asta Õimu "Sünonüümisõnastiku" ja "Antonüümisõnastiku" põhjal koostatud Filosofti arvutitesaurus MS Office'i koosseisus.

1996. aastaks oli selge, et lisaks eesti keele morfoloogia ja süntaksi arvutile arusaadavaks tegemisele on edaspidi vaja ka sõna-semantikal põhinevat leksikaalset andmebaasi. Mujal maailmas teostatud semantiliste arvutileksikonide seas ringi vaadates tundus sobivaim olevat WordNeti idee. Alates 1996. aastast on eesti üldkeele tesauruse loomist toetanud Eesti Teadusfond ja alates 1997 aastast Eesti Informaatikakeskus sihtprogrammis "Eesti keeletehnoloogia" Aastatel 1998–1999 õnnestus osaleda mitmekeelse tesauruse loomisel projekti EuroWordNet-2 raames.

Juba loodud osa eesti keele tesaurusest hõlmab substantiive, verbe ja adjektiive. Tööd alustasime nimisõna- ja tegusõnamõistetest, omadussõnamõisteid on lisama hakatud alles käesoleval aastal ja nende eripärast tuleb pikemalt juttu käesoleva kogumiku eraldi artiklis. Tegijateks on professor Haldur Õimu juhtimisel olnud arvutuslingvistika uurimisrühma liikmed Neeme Kahusk, Leho Paldre, Heili Orav ja Kadri Vider. Tuleb märkida, et kuigi sarnast tööd tehakse EuroWordNeti partnerite juures (pool)automaatselt, oleme meie siiski eelistanud käsitsitööd, ühelt poolt suurema täpsuse huvides, teisalt aga paraku sobivate elektrooniliste lähtetekstide puuduses. Et tesaurus täieneb ja paraneb töö käigus, on artikli näidete ja tabelite aluseks variant 15. maist 2000.

Artiklis antakse lühiülevaade WordNetist ja EuroWordNetist (üksikasjalikumat kirjeldust saab lugeda Kadri Videri ja Heili Orava (1998) artiklist Keeles ja Kirjanduses) ning tutvustatakse eesti keele tesauruse loomist etappide kaupa, peatudes lühidalt ka ilmnenud probleemidel, mida võiks edaspidi lähemalt uurida.

2. WordNet kui arvutitesauruse eeskuju

Semantiline andmebaas on leksikaalse andmebaasi (*lexical database, LDB*) erijuhtum, milles andmebaasi kirjeid, antud juhul leksikaalseid üksusi, seovad omavahel erinevat tüüpi viidad, antud juhul semantilised suhted (Calzolari 1990). Semantiline andmebaas võib leksikaalsete üksuste kohta sisaldada ka süntaktilist ja morfoloogilist infot, ent selle olemasolu pole esmatähtis. Püütakse küll teha vahet keelelisel teadmisel (*linguistic knowledge*) traditsioonilise sõnaraamatu seletuses ja maailma-teadmisel (*world-knowledge*) keeruliste abstraktsete tunnustega kirjeldatud mõistete võrgustikus, kuid rahuldavat määratlust kummagi eristamiseks teisest pole suudetud anda (Bloksma jt 1996). Et eesti keele tesaurus põhineb olemasolevatel traditsioonilistel sõnaraamatutel ja tekstikorpusel (mis annab teavet sõnakasutusest), võib semantilist informatsiooni, mida andmebaas sisaldab, rahumeeli pidada keelelisel teadmisel põhinevaks.

Semantilist andmebaasi võib küll leksikaalsete üksuste kaupa indekseerida ja sortida, kuid peamiseks tunnetuslikuks üksuseks on sellises andmebaasis semantiline väli – semantiliselt seotud sõnade hulk, mis moodustab teatud mõistelise terviku (Õim 1997).

Semantilist andmebaasi, mis keskendub **mõistele** ja **semantiliste suhete** kaudu tema semantilisele väljale, võib nimetada **tesauruseks**.

Tartu Ülikooli arvutuslingvistika uurimisgrupis loodava eesti keele tesauruse põhimõtteliseks eeskujuks on Princetoni Ülikoolis loodud **WordNet** (WN), mida loojad iseloomustavad kui "leksikaalsete viidete süsteemi, mille ülesehitus põhineb psühholingvistilisetel teooriatel inimpsüühika leksikaalsest organisatsioonist ja mälust." (Beckwith jt 1990).

Inimeste teadvuses on sõnad, mõisted ja nendevahelised seosed organiseeritud mentaalsesse leksikoni. 1985. aastal otsustas rühm Princetoni Ülikooli psühholingviste eesotsas George Milleriga luua inglise sõnavara leksikaalne andmebaas, kus sõnad ja tähendused

oleksid organiseeritud samadel põhimõtetel kui mentaalses leksikonis, seega mitte alfabeetiliselt, vaid mõisteliselt. Loodi suur, võimalikult paljusid sõnu ja nende tähendusi haarav leksikaal-semantiline andmebaas, mille abil taheti kontrollida leksikoloogiasse kalduvaid psühholingvistilisi hüpoteese. Seni oli leksikaalset esindust puudutavaid psühholingvistilisi hüpoteese kontrollitud piiratud hulga – kuni sajakonna – enamasti nimisõnalise näite peal. Miller koos oma kaastöötajatega tunnistab, et ainult osa neist hüpoteesidest on kinnitust leidnud, võiks öelda, et kõige tugevamad jäid ellu. Selle suure töö peamine psühholingvistiline tulemus on väide, et süntaktiliste kategooriate (sõnaliikide) fundamentaalsed erinevused on selgelt eristatavad ja kasutusel ka nende semantilises organisatsioonis. Nii on substantiivid leksikaalses mälus organiseeritud temaatiliste hierarhiatena, paljud adjektiivid kui n-mõõtmelised klastrid ja verbid on omavahel mitmekesistes osalussuhetes (Miller jt 1990). Leksikaalse andmebaasi mõistelise ülesehituse põhimõte piiras ka käsitletavate sõnade hulka; andmebaasi märksõnadeks said olla vaid täistähenduslikud sõnad. Samuti valiti põhimõtteliselt lõputust semantiliste suhete hulgast välja hulk olulisemaid, mis võeti kasutusele.

3. EuroWordNet – mitmekeelne leksikaal-semantiline andmebaas

EuroWordNet (EWN) oli Euroopa Komisjoni projekt aastatel 1996–1999, mille eesmärgiks oli luua WNi eeskujul mitmekeelne leksikaal-semantiline andmebaas, milles erinevate keelte (inglise, hollandi, itaalia, hispaania, prantsuse, saksa, tšehhi, eesti) wordnetid on ühendatud keeltevahelise indeksi kaudu.

EWN peamine erinevus WN-st on tema mitmekeelsus. Kõik projektis osalejad lõid WNi põhimõttelisele ülesehitusele toetudes omakeelse wordneti, kuid keeltevahelise indeksi (*interlingual index*, ILI) kaudu on võimalik leida sama mõistet väljendavad sünonüümihulgad teistes keeltes. EWN täisandmebaas sisaldab seega 2 erinevat tüüpi mooduleid:

1. Keelest sõltuvad moodulid (iga keele wordnet eraldi), mis omakorda jagunevad:
   1.1. sõna – tähenduse kirje (*word-meaning*, W-M);
   1.2. sõna – üksikobjekti kirje (*word-instance*, W-I);

1.3. keelesisesed suhted – semantilised suhted ühe keele mõistete vahel;
1.4. keeltevahelised suhted – semantilised suhted ILI-kirje ja konkreetse keele mõiste vahel.

2. Keelest sõltumatud moodulid, milleks on:
2.1. keeltevahelise indeksi (ILI) moodul;
2.2. valdkondade (*domain*, DOM) ontoloogiate moodulid;
2.3. tippmõistete (*top concepts*, T-C) ontoloogia moodul.

Kokkuleppeliselt on keeltevaheline indeks ingliskeelne ja koosneb suures osas WN kirjetest. Et aga leksikaalsed tähendusväljad erinevates keeltes võivad olla nihkes või teisiti allmõisteteks jaotatud, nii et tekivad "leksikaalsed augud", ei ole ILI-kirjete suhted keele kirjetega alati üksüheselt sünonüümsed.

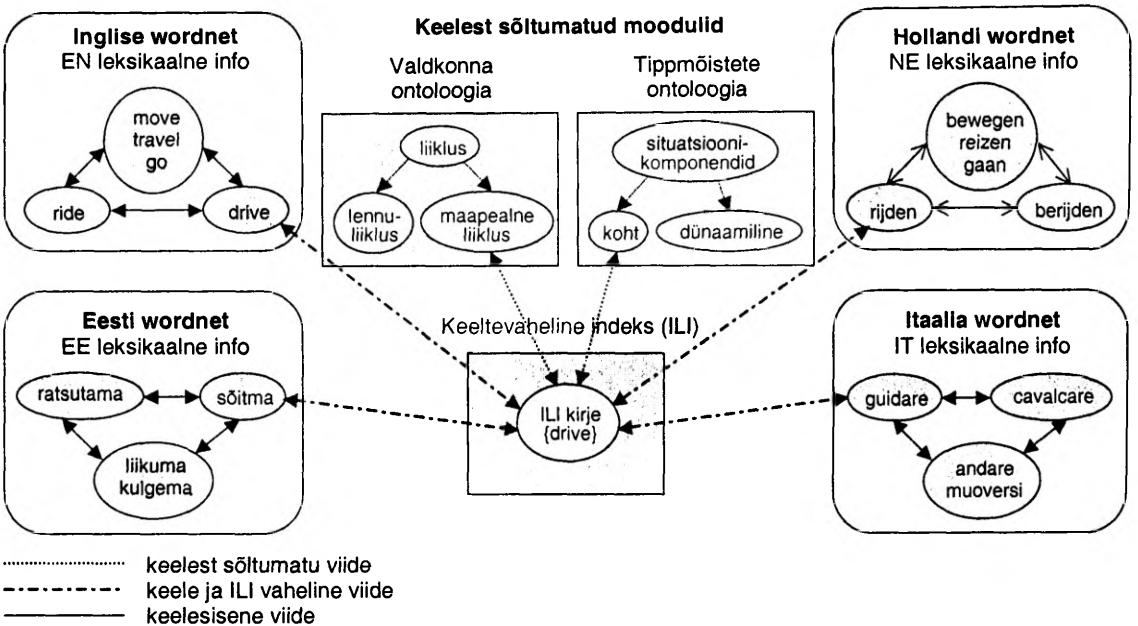
Uus moodul võrreldes WNga on ka **tippmõistete ontoloogia** (WN esitab küll hierarhiate tipud ehk unikaalsed alustajad), milles on püütud formuleerida keelest sõltumatut tegelikkuse põhilist semantilist jaotust (Bloksma jt 1996: 55), selle loomisel on tuginetud ka teistele samalaadsetele ontoloogiatele. Ontoloogia püüab anda kõige üldisemat mõistelist jaotust, mis on ühine kõigi osalevate keelte wordnet'ides. Kui EWN alustamisel olid tippmõisted ka omavahel hierarhilistes seostes, siis praegu on sellest loobutud ja tippmõisteid käsitletakse üha enam pigem **semantiliste tunnusjoontena** (*semantic feature*), mille erinevaid kombinatsioone võib omistada igale mõistele.

Leksikaalse informatsiooni sisestamiseks andmebaasi, selle töötlemiseks ja semantiliste suhete loomiseks on kasutatud peamiselt Lernout&Hauspie loodud sisestusliidest Polaris. Eesti keele tesauruse andmebaas eksisteeribki EWN andmebaasina (keelest sõltuv moodul) Polarise formaadis.

4. Põhimõisted

Selguse huvides oleks tarvis lahti seletada mõned terminoloogilised seisukohad kasutatavas paradigmas.

Keelekasutaja teadvuses kujunenud **mõistel** (*concept*) on oma kindel kontseptuaalne **tähendus** (*meaning, sense*), mida väljendatakse kasutuses samatähenduslike ehk sünonüümsete **sõnadega** (laiendatult: leksikaalsete üksustega) või **seletuses** mingil muul viisil semantiliselt (so. tähenduslikult) seotud sõnade kaudu. Samas,



Joonis 1. Wordneti moodulid (eestindatud EWN lõppdokumendist, Diez-Orzas jt 1996)

keelekasutuses eksisteeriv sõna ise võib olla nii **ühetähenduslik** ehk monoseemiline kui ka **mitmetähenduslik** ehk polüseemiline. Mõiste, mida saab väljendada üldtuntud sõnaga või pikema kinnistunud leksikaalse üksusega, on keeles **leksikaliseerunud**; mõiste, millest keelekasutaja pikema seletuseta aru saada ei suuda, on tema teadvuses **leksikaliseerumata**. Ka **mõistepiirid** sõltuvad igaühe individuaalsest keelelisest kogemusest.

Wordneti elementaarosake on sünonüümirida – **sünohulk** (*synonym set, synset*), mille moodustavad ühte mõistet väljendavad sünonüümsed sõnad ja sõnaühendid. Termini sünohulk oleme loonud sellepärast, et erinevalt sünonüümisõnastiku sünonüümireast võib meie sünohulk olla ka üheliikmeline. Kui sünonüümisõnastiku eesmärgiks on kõigi võimalike keeles leiduvate sünonüümide esitamine, siis meie töö eesmärgiks on mõistete esitamine, ka siis, kui selle väljendamiseks keeles leidub ainult üks leksikaalne üksus. Wordnet-tüüpi tesauruses eksisteerivad **semantilised suhted** seega sünohulkadena esitatud mõistete vahel ning tundub loomulik pidada semantilisi suhteid sünohulkade vahelisteks erinevat tüüpi viitadeks.

5. Eesti keele arvutitesauruse koostamise allikad

Peamist leksikaalset informatsiooni teatud keele tesauruse koostamiseks annavad loomulikult **ükskeelsed** tähendusi eristavad ja/või seletavad või sünonüüme esitavad sõnastikud. Eesti keele kohta pole selliseid palju, veel vähem leidub neid, millest on kättesaadav arvutiversioon või elektrooniline tekstivariant:

- **Eesti Kirjakeele Seletussõnaraamatu** (EKSS) elektrooniline tekst on saadud koostajatelt Eesti Keele Instituudist, teisendused lihtsasse tekstiandmebaasi on teinud Leho Paldre;
- **KeeleWebi** [http://ee.www.ee/] ühispäringu kaudu oleme kasutanud Filosofti tesaurust, Asta Õimu Sünonüümisõnastikku, Antonüümisõnastikku ja Fraseoloogiasõnaraamatut.

Mitme allika kasutamine elimineerib idiosünkraasia ja ühe allika omapära. Olulised leksikaal-semantilised jooned saavad tõestatud esinemisega paljudes allikates.

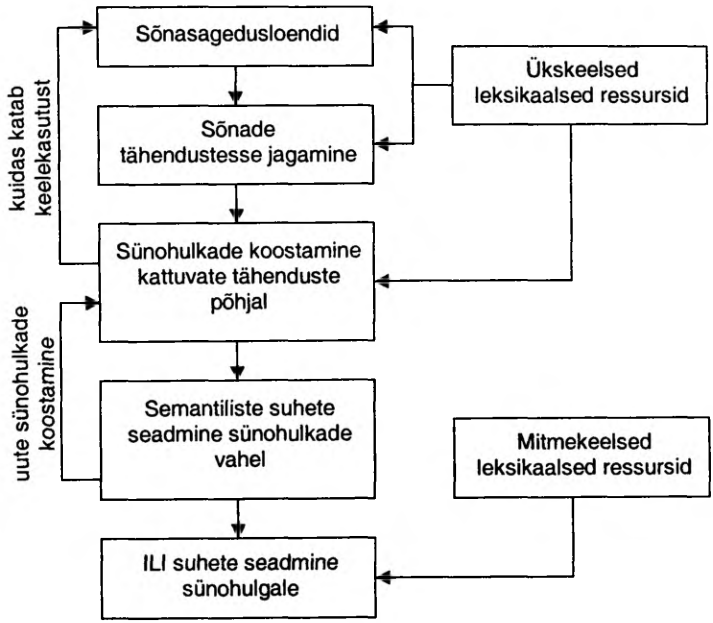
Mitmekeelseid allikaid on vaja läinud tesauruse eestikeelse osa sidumisel keeltevahelise indeksiga (ILIga):

- J. Silveti Inglise–eesti sõnaraamat, TEA Password;
- P. Saagpaku Eesti–inglise sõnaraamat.

Kuna ILI on ingliskeelne, tuli mõnikord appi võtta ka inglise keele seletavad sõnaraamatud nagu Webster, COBUILD-Collins jt

6. Koostamise käik

Töö eesti keele arvutitesauruse koostamisel on ainulaadne, seetõttu peab tutvustama protsessi, mis siiani läbitud, veidi lähemalt. Iga sammu juures tuleb ka juttu kerkinud probleemidest ja käesoleva hetke tulemustest selles aja- ja mõttemahukas töös. Koostamise peamist skeemi illustreerib joonis 2.



Joonis 2. Eesti keele arvutitesauruse koostamise käik

6.1. Baasmõisted ja sõnasagedusloendid

Mõistelisel alusel organiseeritud tesaurust oleks ilmselt mõistlik hakata koostama mingitest üldisematest sõnadest või mõistetest lähtudes. Nii oli ka eesti keele wordnet-tüüpi tesauruse loomisel töö esimeseks etapiks **baasmõistete** *(base concepts)* leidmine ja kirjeldamine. Baasmõisted moodustavad keele wordneti tuuma ning esindavad keele peamist semantilist jaotust. Baasmõisteid eristavad teisteist mõistetest:

a) suur hulk hüponüüme ehk alammõisteid;
d) hüperonüümideks ehk ülemmõisteteks on tippmõisted või on nad hierarhias tippmõistetele üsna lähedal.

Milleri ja Fellbaumi (1991) ning EuroWordNeti (Vossen jt 1998) määratluste järgi esitavad baasmõisteid sõnad,
a) mis on suure esinemissagedusega nii tekstides kui sõnaseletustes;
b) mida on raske defineerida või seletada, kasutatakse sageli ringseletamist;
c) millel on palju ja sageli raskesti eristatavaid tähendusi (nt sõnal *asi* on tesauruses 11 tähendust, sõnadel *minema* ja *ajama* 17 tähendust);
d) mille tähendusi eristab tihti lai valik süntaktilisi jooni.

Eesti Kirjakeele Korpusest ja EKSSi seletustest korjatud **sõnasagedusloendid** vastavad nii baasmõistete moodustamise kriteeriumile kui annavad ka aluse jälgida, et sagedasemate sõnadega esitatud mõisted oleks tesauruses kõigepealt olemas.

Puudusena tuleb siiski nentida, et verbimõistete moodustamiseks on nõnda saadud vaid lihtverbide loend. Olemasolevad keeletehnoloogilised vahendid ei suuda veel kahjuks ühend- ja väljendverbe omaette üksustena lemmatiseerida. Liitverbide eristamise korral asuksid mõnedki verbid (nt *tooma*) hoopis sagedaste liitverbide moodustajate eesotsas (Vider 1997). Et puudub ühend- või väljendverbide loend, ei analüüsi ka olemasolevad morfoloogia- ja süntaksianalüsaatorid neid ühe tähendusüksusena ega suuda eristada.

Samal ajal on baasmõisted kui ülemine osa mõistelisest hierarhiast piisavalt üldised. Mõnikord tekib seetõttu tarvidus lisada tesaurusse või luua juurde suurt hulka hüponüüme kokkuvõtvaid terminilaadseid sõnaühendeid, mis üldkeele kasutuses leksikaliseerunud pole, nagu näiteks sünohulk *'psüühilised nähtused'* Tekib vastuolu mõistelise kategoriseerimise ja keelelise väljenduse vahel, kuid taksonoomia terviklikkuse huvides tuleb minna kompromissile leksikaliseerituse nõudega.

EWN projekti käigus koostati keelte wordnet'e peamiselt kahel viisil (Vossen 1998):
1) loodi oma keele sobivate leksikaal-semantiliste allikate põhjal (pool)automaatselt vastava keele sünohulgad ja otsiti neile ILI hulgast lähimad vasted;

2) adapteeriti või tõlgiti WN sünohulgad oma keelde, nõnda oli vastavus ILI kirjetega üksühene. See meetod ei too aga välja keelespetsiifilisi leksikaal-semantilisi nähtusi ning võib esitada koguni ühekülgse pildi vastava keele mõistelisest jaotusest.

Eesti keele tesauruse tegemisel on kasutatud peamiselt esimest viisi. Järgides aga EWN projekti nõuet leida üksühene vastavus 1310 nn ühise baasmõistega (*common base concepts*) kõigis kaheksas osalevas keeles, tuli kasutada ka teist meetodit.

Lühidalt kokku võttes seisnes tesauruse loomise I etapp baasmõistete koostamises:

1) EKSSi seletuste sõnasagedusloendist saadi 1133 substantiivi (sagedusega F>46) ja 516 verbi (sagedusega F>40);
2) neid loendeid võrreldi tõlgitud baasmõistete sõnaloenditega, et selekteerida välja juba olemas olevad sagedasemate sõnade tähendused, eeldatavalt baasmõisted;
3) puuduolevate sõnadega koostati sünohulgad, mis peaks esitama nende sõnade kõige üldisemaid, baasmõistete alla kuuluvaid tähendusi;
4) kõige eelneva tulemusena saadi 698 substantiivi sünohulka ja 366 verbi sünohulka, mis võiks moodustada eesti keele baasmõistete osa.

Järgmises etapis, tesauruse laiendamisel baasmõistetest kaugemale, oli uute sünohulkade moodustamise üheks aluseks samuti sõnasagedusloendid. Sedapuhku võtsime arvesse sõna esinemissagedust ka eesti keele tekstikorpuses.

6.2. Tähendustesse jagamine ja sünohulkade koostamine

Sünohulga tasandil pole loodavas wordnet-tüüpi eesti keele tesauruses palju uut, peamiselt on tegemist seletussõnaraamatu ja sünonüümisõnastike **leksikaalse informatsiooni ümberstruktureerimisega**.

1. Jagades EKSSi märksõnad tähendusteks, saame täpselt sama palju sünohulki, kui palju on sõnatähendusi seletussõnaraamatus. Ometi pole isegi tähendustesse jagamine sugugi lihtne. EKSSis avatakse sõna tähendus kas (a) lühikese kirjeldava seletusena, mis on filoloogilist laadi, kuid võib sisaldada ka entsüklopeedilisi elemente, (b) sünonüümide kaudu, (c) tulenemisseletuse kaudu või (d) antonüümi kaudu:

(koostatud EKSS kirjete põhjal)
jääma 3. muutuma, saama. a. (kellekski, millekski v mingisuguseks). Jäi leseks, vaeslapseks...
minema 5. Oma seisundit, olekut v. asendit muutma; senisest erinevaks muutuma. a. (elusolendiga toimuvate füüsiliste, füsioloogiliste, psüühiliste jm muutuste kohta). Laps on ulakaks, ülekäte läinud...
muutuma 1. teistsuguseks või täiesti teiseks saama. Ta muutus näost kahvatuks...

2. Liites kokku suurel määral kattuvad sünonüümsed tähendused, peaksime saama sünohulgad, mis esitavad sünonüümide ja seletuste kaudu üksteisest erinevaid mõisteid. See on punkt, kus ilmneb, kui vajalik on erinevate leksikaalsete ressursside elektrooniliste versioonide olemasolu ka tulevaste leksikoloogiliste-semantiliste tööde tarvis. Kui on olemas EKSSi elektrooniline tekst, on võimalik üles leida ja läbi vaadata kõik sarnased tähendusseletused kogu sõnastikus. Eelnevast näitest saab sel juhul tuletada uue näite:

sünohulk = **jääma 3, minema 5, muutuma 1, saama 1** – kellekski, millekski või mingisuguseks, senisest erinevaks, teistsuguseks või täiesti teiseks muutuma. Jäi leseks, läks hulluks, muutus kahvatuks, sai terveks...

Arvutiga leksikoniloomise ajastul oleks infootsingu seisukohalt hea, kui sõnaseletused või definitsioonid järgiksid mingeid tüüpilisemaid formalisme. Sõnaseletus või definitsioon peaks sisaldama piisava hulga semantilisi tunnuseid, mille järgi saaks moodustada (ka automaatselt) võimalikult palju erineva semantilise iseloomuga viiteid teistele mõistetele. Näiteks Pisa Arvutuslingvistika Instituudi juures loodav LDB püüab loomuliku keele sõnade definitsioone esitada hüperonüümide ja hüponüümide kaudu, et mõistete süsteeme saaks kujundada automaatselt (Calzolari 1990).

Oleme oma töös eesti keele tesauruse kallal püüdnud juhinduda mõttest, et seletused tuleb allikast valida või ise koostada sellised, et nad kehtiksid kõigi sünohulga liikmete kohta. Lubamatu on olukord, kui sünohulga liikmete järel olevad seletused on vastuolulised või teineteist välistavad. Tuleb püüda vältida ka olukorda, kui ühe sünohulga liikme seletamiseks kasutatakse mõnd teist sünohulga liiget. Selguse mõttes on mõnikord hea kasutada seletuses vastandust.

Esineb ka seletusi, mis koosnevad püsiühendeist, nt sünohulgas *enamik1*, *enamus1* – 'suurem osa' võiks ka 'suurem osa' olla sünohulga liikmeks, mitte seletuseks. Täpsed kriteeriumid sünohulga liikme määratlemiseks siiski puuduvad, peamiseks oluliseks ja samas

ähmastavaks tunnuseks on sünohulga liikme esinemine kasutuses mõistelise üksusena.

Omaette probleem on tähenduste eristamise motiveeritus juba allikateski. Esineb tähenduste **üle-eristamist**, milles kahe tähenduse vaheline erinevus on väga ähmane (sageli tegelikult vaid kasutuskontekstist sõltuv), kui ka **ala-eristamist** (seda maksab kahtlustada, kui seletuses esineb osalauseid rinnastav *või*, mis viitab tegelikult kahele erinevale tähendusele).

6.3. Sünohulga seostamine erinevat tüüpi semantiliste suhete kaudu teiste sünohulkadega

Erinevaid semantilisi suhteid keeles ei saa määratleda lõpliku hulgana, semantilise andmebaasi jaoks aga on vältimatu panna paika mingi lõplik loend. EWN loojad on oma loendi koostamisel tuginenud paljudele semantilistele teooriatele – alates Roget' tesauruse mõistekäsitlusest, läbi kognitiivse ja leksikaalse semantika kuni teadmiste formaalsete esitussüsteemideni välja (Bloksma jt 1996). Lõpliku valiku tegemise juures on välja töötatud ka testid nende suhete kindlakstegemiseks konkreetsete sõnade vahel (Orav 1998).

Kõikide semantiliste suhete seadmisel rakendatakse **ühtsusprintsiipi**: kahe sünohulga vahel saab kehtestada üht ja ainult üht tüüpi semantilise suhte, st nt *mööbel* ja *tool* ei saa olla korraga alluvussuhtes ja osa-terviku suhtes (Alonge jt 1998). Automaatselt on kõik suhted kahesuunalised (kuigi mitte kõik suhted pole sümmeetrilised): kui luuakse semantilise suhte viit ühelt sünohulgalt teisele, lisab sisestusliides viida ka teiselt sünohulgalt esimesele. Sümmeetriliste suhete nagu sünonüümia ja antonüümia korral on viidad ühe nimega, asümmeetriliste suhete viidad on aga erinevate nimedega, nt *causes* ja *is_caused_by*.

Enamasti luuakse EWN andmebaasis semantilisi suhteid ühe ja sama sõnaliigi kirjete vahel, mõned suhted (nt osalussuhe/rollisuhe) on võimalikud ainult kindlate sõnaliikide vahel. Kui semantiline suhe võib esineda üle sõnaliigi piiride, märgistatakse seda laiendiga *_xpos_ (x_part_of_speech)* viida sees.

Johtuvalt EuroWordNeti prioriteetidest on seni ka eesti keele tesauruses keskendutud enim sünonüümiale kui fundamentaalsele semantilisele suhtele ja hüperonüümia/hüponüümia suhtele kui olulisi semantilisi hierarhiaid loovale suhtele.

Tabel 1 esitab ülevaatlikult eesti keele tesauruse arvulise seisu sünohulkade, sõnatähenduste, sõnade ja sõnaühendite ning semantiliste suhete esindatuse vaatepunktist.

Tabel 1. Eesti keele tesaurus

	Nimisõnades	Verbides	Adjektiivides	Kokku
Sünohulki	6236	2650	307	9193
Sõnade tähendusi (variante)	10463	5616	518	16597
Tähendusi sünohulga kohta	1,68	2,12	1,69	1,8
Erinevaid sõnu ja sõnaühendeid	8928	3755	419	13102
Tähendusi sõna kohta	1,17	1,5	1,24	1,27
Semantilisi suhteid	13361	5316	538	19215
Semantilisi suhteid sünohulga kohta	2,14	2,0	1,75	2,09

Allpool vaatleme lähemalt olulisemaid semantilisi suhteid nii noomenite kui verbide vahel.

6.3.1. Sünonüümia

Lähtudes sünohulga definitsioonist, on sünonüümiasuhe fundamentaalne. Üldlevinud definitsioon sünonüümiale ütleb, et sõnad on sünonüümsed, kui üht sõna saab lauses asendada teisega ja lause tähendus jääb samaks. Selle definitsiooni järgi tõeseid sünonüüme leiab väga harva, erinevused on tingitud peamiselt stiili, kuid ka tundesisu ja tahterõhu varjundeist. Cruse (1986) sõnatähenduse käsitlus määratleb: kaks leksikaalset üksust on absoluutsed sünonüümid (st. neil on identsed tähendused) siis ja ainult siis, kui nende kontekstuaalsed tähendused on identsed. Loomulikult on võimatu kontrollida väite paikapidavust kõikides kontekstides.

Nõrgendatud versioon sellele definitsioonile kõlab: kaks sõna on sünonüümid kontekstis C, kui ühe asendus teisega kontekstis C ei muuda lause tõeväärtust. Leksikaalse semantika teoorias piisab semantilisest sarnasusest ja suhe peab olema ümberpööratav: kui x on sarnane y-le, siis y on sama sarnane x-le. Asta Õim lähtub oma sünonüümikäsitluses "sõnast vm keeleüksusest kui tekstilise

kasutuse üldistusest vastandina senistele käsitlustele, mis on lähtunud sõnast kui objektiivset maailma peegeldavast sõnavaralisest üksusest." (Õim 1991: 4). Sünonüümiasuhte määratlemiseks kasutab ta kolme tunnust:

L – keeleüksuste leksikaalse tähenduse samasus;
G – grammatilise sõnaliigilise kuuluvuse samasus;
S – süntaktilise rolli samasus lauses.

Kõigi kolme tunnuse olemasolu korral võib keeleüksusi (nii sõnu kui sõnaühendeid) pidada tsentraalseteks sünonüümideks, ainult kahe tunnuse (üks neist L) kehtivus lubab keeleüksusi pidada perifeerseteks sünonüümideks. Nii lubab ühiste tunnuste L ja S olemasolu lugeda ühend- ja väljendverbe lihtverbide sünonüümideks, nt *alustama, peale hakkama, pihta hakkama*. Tuleb siiski möönda, et tunnus L ehk leksikaalsete tähenduste samasus võib olla kaunis subjektiivne ja sõltuda individuaalsest keeletajust. Sünonüümisõnastikus on tegemist dominandi ehk põhisõna ümber koondunud sünonüümiridadega, mille äärmised liikmed meie töö mõistes täissünonüümid ei pruugi olla ja seega ka ühte sünohulka kuuluda ei saa.

Kontekstis kasutatakse sünonüümsetena ka sõnade üldisemaid tähendusi ehk hüperonüüme, mistõttu tuleb meie töö lahendustes püüda vältida sõna x spetsiifilisemate tähenduste sattumist sama sõna kõige üldisema tähenduse alla. Sellist, tegelikult tähenduse ülieristamise juhtumit on EWN töödes nimetatud ka pragmaatiliseks eristuseks (Vossen jt 1998):

sünohulk(x1,y1,z1)
=> sünohulk(x2,b2)
=> sünohulk(c2,z2)

Vaatleme kahte näidet:

kriis1 – 'raske, terav, komplitseeritud olukord, ohtlik, vastuoludest lõhestatud seisund'
=> kriis4, kitsikus1, nappus1 – 'millegi puudus' "Sõjaajal andis tunda toiduainete kriis"

Antud näites on lahenduseks *kriis4* kustutamine sünohulgast, sest kuigi näitelauses on kasutatud *kitsikuse* või *nappuse* asemel *kriisi*, ei ole see *kriisi* omaette tähendus, vaid vastab täielikult *kriis1* üldisemale tähendusele.

Mõnes lahenduses on kitsenduse eiramine siiski vältimatu:

mees2, meessoost inimene1, meesterahvas1 – 'sellest soost inimene, kes ei saa sünnitada lapsi';
=> mees1 – 'täiskasvanud meessoost inimene, vastandina poisile';
=> mees4, abielumees1

Et sünonüümiasuhe ongi see, mis sünohulga liikmeid omavahel üheks mõistetervikuks seob, siis täissünonüümiat eraldi suhtega ei märgistata. Mittetäpseteks ehk kvaasisünonüümideks tuleb nimetada sõnu, mille seletused langevad kokku ainult osaliselt, erinedes teineteisest mõne vähemolulise komponendi poolest (Hallik 1990). Lähisünonüümiat tähistatakse EWN andmebaasis viitadega

leidma2, avastama2, märkama3 **near_synonym** silmama1, märkama2, tähele panema2
liikuma2, siirduma2, kulgema2 **xpos_near synonym** kulgemine1

Regulaarseid küllalt püsiva tähendusega tuletusliiteid saab ära kasutada (pool)automaatselt semantiliste suhete seadmisel sünohulkade vahel. Absoluutse produktiivsusega *mine*-sufiks ei lisa tuletusalusele verbile mingeid semantilisi tunnuseid, vaid muudab ainult sõnaliiki. Tuletus on puhtalt süntaktiline ja *mine*-liitelised substantiivid väljendavad tegevust kõige üldisemas mõttes, kui protsessi või nähtust (Kasik 1996). Seega on küllaga põhjust seada sõnasagedusloendite põhjal tesaurusse tulnud *mine*-tuletistele *xpos_near_synonym* suhet ja ka vastupidi. Ent *mine*-tuletised väärivad muulgi põhjusel tähelepanu – teatavasti esinevad muidu lahkukirjutatavad ühend- ja väljendverbid *mine*-tuletistena kokkukirjutatuna ja nende sage esinemine annab põhjuse ka vastavad ühend- ja väljendverbid tesaurusse tuua.

6.3.2. Hüperonüümia/hüponüümia

Semantiline suhe, kus üks klass sisaldab teist, on hüperonüümia/hüponüümia suhe. Hüponüümiaga tähistatakse tähenduste hierarhilisi alistussuhteid. Alammõiste sõna on oma ülemmõiste suhtes hüponüüm, ülemmõiste sõna on oma alammõiste suhtes hüperonüüm. Ülemmõiste ja tema alammõisted moodustavad hierarhilise puu, milles ülalpool on väga suuremahulised üldised mõisted, allapoole eritunnuste hulk kasvab ja mõistete maht väheneb.

EWN andmebaasis kasutatakse hüpo- ja hüperonüümiasuhte viitadeks:

muutuma1, teisenema1 **has_hyponym** kasvama4, suurenema3
kasvama4, suurenema3 **has_hyperonym** muutuma1, teisenema1
liikumine3 **has_xpos_hyponym** liikuma2, siirduma2, kulgema2
liikuma2, siirduma2, kulgema2 **has_xpos_hyperonym** liikumine3

Hüperonüümi leida pole keeruline, sest sageli seletatakse sõnu üldisema kaudu, kuid hüperonüümiasuhete seadmisel on oluline järgida **ökonoomiaprintsiipi**: kui M1 hüperonüümiks on M2 ja M2 hüperonüümiks on M3, siis saab M1 hüperonüümiasuhte M2-ga kui vahetu hüperonüümiga, aga mitte M3-ga:

Kontrabass – 'suurim ja madalaima häälega keelpill'
hüperonüüm => poogenpill, vibupill – 'keelpill, millel heli tekitatakse poognaga'
hüperonüüm => keelpill

Hüpo-/hüperonüümia verbileksikonis pole siiski päris sarnane noomenite omale, nt WN-s on seda nimetatud troponüümiaks, mis näitab, mil *viisil* on üks verb spetsiifilisem ja tähenduse poolest kitsam kui teine (Beckwith jt 1990).

Ideaalis peaks igal mõistel olema üks hüperonüüm ja selle poole püüdlesime ka eesti keele tesauruse tegemisel. Ometi, erineval liigitusalusel võib üks ja seesama mõiste hierarhiate täiuslikkuse huvides kuuluda mitmesse hierarhiasse. Tüüpilisemaid näiteid sellistel juhtudel on mõiste kuulumine erialasesse taksonoomiasse (hobunel *has_hyperonym* hobuslane1) ja üldkeelsesse liigitusse (hobune1 *has_hyperonym* koduloom1). Pole ka mingit mõtet luua kummagi hüperonüümi jaoks eraldi sünohulka, sest tegu on ikkagi sama tähendusega.

Pärast baasmõistete hüpero-/hüponüümiasuhete seadmist ja sellega üldiste hierarhiate väljakujundamist oli tesauruse loomise II ja III etapi tööd seotud nende hierarhiate edasiarendamisega järgmis(t)e taseme(te) hüponüümide kaudu. Ka kõigile sõnasagedusloendite põhjal moodustatud sünohulkadele tuli leida vähemalt hüperonüüm. Nõnda oli EWN projekti lõppedes eesti wordnetis 82 verbimõistete hierarhia tippu ja 25 noomenimõistete hierarhia tippu (kooskõlas Princetoni psühholingvistide väitega sõnaliikide semantilise struktuuri erinevusest) ning hüperonüümiasuhte esinemine sünohulkade vahel absoluutselt regulaarne. Hierarhiate sügavus oli

üsna erinev, kõige sügavam oli 12 hüponüümide tasandit. Kõige rohkem hüponüüme paiknes 3.–6. tasandil.

Tabel 2. Hierarhiate tipud ja nende hüponüümide hulk

Nimisõnamõisteid	Hüpon	Verbimõisteid	Hüpon
Olev2	3011	sooritama4, teostama4, tegema5	952
Abstraktsioon2, üldmõiste1	1645	muutma2, transformeerima1	369
Tegevus1, tegutsemine1, toiming2, toimetus1	432	muutuma1, teisenema1	294
Psüühiline nähtus1	243	liikuma3	223
Tegu3, toiming3	330	tegema6, tekitama3, põhjustama1, tingima3	208
Fenomen1, ilming2, nähtus1, nähtumus1	298	olema8	122
Grupp1, rühm2, hulk4	285	olema4, eksisteerima2, olemas olema1, olelema2	76
Koht4, asukoht1, paik1, asupaik1, asend1, paiknemine1	204	looma2	54
Sündmus2	125	otsustama3, hindama2, hinnangut andma1	46
Osa2, tükk1	111	mõtlema1	42

Järgnevalt vaadeldakse semantilisi suhteid, mille esinemine eesti keele tesauruses on töö praeguses etapis veel täiesti ebaregulaarne, et mitte öelda juhuslik.

6.3.3. Antonüümia

Antonüümia on tähenduste vastandlikkus, täpsemalt: vastandussuhe ühe semantilise tunnusjoone, ühe tähenduskomponendi alusel, seega nende muu tähendussisu langeb kokku. Näiteks *mees* ja *naine* on mõlemad *inimese* hüponüümid, aga *mees* on *naise* vastand. On ka võimalik, et sõnade hüperonüümid on omavahel antonüümid (*müüma* ja *ostma* on *andma* ja *võtma* hüponüümid). Milleri jt (1990) artiklis käsitletakse antonüümiat kui leksikaalset suhet – kahe erineva sünohulga kõik liikmed ei pruugi omavahel antonüümid olla. Nt moodustab ühe sünohulga *tarbija1, kasutaja1, pruukija1*. Sõna *tootja* võib olla küll *tarbija* antonüüm, kuid vastandus *pruukija:tootja* ei kõla just eriti veenvalt. Võimalik, et antonüümiasuhte

puhul on tegemist millegi enama kui tähenduste vastandlikkusega ja et see vajaks tervet eraldi uurimust. Adjektiivide puhul on antonüümiasuhe eriti oluline, sellest on täpsemalt juttu Heili Orava artiklis käesolevas kogumikus.

Praegu võib piirduda sellega, et ka Asta Õim nimetab antonüüme leksikaalsemantilise paradigma liikmeteks (Õim 1995) ja et EuroWordNetis on antonüümiasuhe küll sünohulkade vahel, kuid seda pole võimalik määrata täpsustamata milliste sünohulga liikmete vahel see suhe kehtib:

lõpetama2, lõpule viima1 **antonym** algama4, alustama2, pihta hakkama2, algust tegema2, peale hakkama3

Lähiantonüümia puhul sellist piirangut ei ole, see määratakse sünohulkade vahel:

meelde tuletama1, meenutama2 **near_antonym** unustama1
vastama1, vastust andma1 **xpos_near_antonym** küsimus3

6.3.4. Osa–tervikusuhted

Semantiline suhe leksikaalsete üksuste vahel, mis näitab tervikut ja selle osasid, on holonüümia/meronüümia. Maailmas on kaks peamist viisi asjade nimetamiseks: leida igale asjale oma nimi või kirjeldada täpselt, millisest materjalist on asi tehtud, kus ta paikneb, millised on tema komponendid jne. Meronüümiasuhet seatakse ainult noomenimõistete vahel ning peale kõige üldisema meronüümiasuhte on EWNs kehtestatud veel mõned alaliigid:

üksus1, kogu3, tervik2 **has_mero_part** eksemplar1, üksikese1
eksemplar1, üksikese1 **has_holo_part** üksus1, kogu3, tervik2
teater1 **has_mero_location** lava3, estraad1
lava3, estraad1 **has_holo_location** teater1
süsi1 **has_mero_madeof** süsinik1
süsinik1 **has_holo_madeof** süsi1
arv2 **has_mero_member** number1
number1 **has_holo_member** arv2
ajutegevus2, mõtlemine1, vaimne tegevus1 **has_mero_portion** mõte1
mõte1 **has_holo_portion** ajutegevus2, mõtlemine1, vaimne tegevus1

Oluline on see, et kõikvõimalikud osad omistataks kõige üldisemale neid osi omavale mõistele. Pole mõtet viidata, et *saba* on osaks nii *koerale*, *kassile* kui *kalale*, tegelikult seatakse ainult suhe 'loom1, elajas1 *has_meronym* saba1' sabatute loomade juurde on võimalik eraldi märkida saba puudumist.

6.3.5. Osalus–rollisuhted

Sõnastikes leidub verbide suhetest teiste sõnaliikidega palju lisainformatsiooni, mis annab lisateadmisi ka sellise tesauruse kasutajale, lisaks on keeled rikkad tuletuste poolest, mis genereerivad nimisõnadest verbe ja vastupidi, nagu nt *jooksma/jooksja, telefon/telefoneerima* jms. See tugev semantiline suhe leksikaalsete üksuste vahel on osalus-/rollisuhe (*involved/role*), mis osutab suhetele verbide ja nimisõnade vahel (mõnel juhul ka adjektiivide ja adverbide vahel) ja mille tähendus on seotud verbi tähendusega. Osalussuhte alaliigid on järgmised:

- agendi-osalus
  rääkima3, kõnelema3, ütlema4 **involved_agent** kõneleja1
  kõneleja1 **role_agent** rääkima3, kõnelema3, ütlema4

- patsiendi-osalus
  rääkima3, kõnelema3, ütlema4 **involved_patient** kuulaja1
  kuulaja1 **role_patient** rääkima3, kõnelema3, ütlema4

- instrumendi-osalus
  rääkima3, kõnelema3, ütlema4 **involved_instrument** jutt1
  jutt1 **role_instrument** rääkima3, kõnelema3, ütlema4

- koha-osalus
  olema7, asetsema1, asuma2, paiknema1 **involved_location** koht4
  koht4 **role_location** olema7, asetsema1, asuma2, paiknema1

- suuna-osalus
  minema15, viima1 **involved_direction** koht4
  koht4 **role_direction** minema15, viima1

Andmebaasi on lülitatud ainult "tugeva osalusega" seotud mõisted. Nt verb *liikuma* lubaks nii patsiendi kui agendi osalust, kuid liikuda võib palju "objekte" ja nad pole väga selgelt seotud selle verbi tähendusega. Põhjalikum freimisemantika alane uurimistöö eesti verbide kohta võib lisada relevantseid seoseid, seda tõi esile ka Heili Orava magistritöö direktiivsetest verbidest (Orav 1998).

6.3.6. Põhjussuhe

Põhjussuhe osutab suhetele verbide vahel, kus tegevus põhjustab mingi sündmuse, protsessi või seisundi (nt *näitama/nägema; andma/omama*). Siin eristatakse põhjuslikkust kolme ajalise suhtega verbide (ka teiste sõnaliikide) vahel:

- põhjuslikkuse suhe kahe situatsiooni vahel, mis on ajaliselt eristatud (*tulistama/(märki)tabama*);
- põhjuslikkuse suhe kahe situatsiooni vahel, mis on ajaliselt osaliselt kattuvad (*õpetama/õppima*);
- põhjuslikkuse suhe kahe situatsiooni vahel, mis on ajaliselt ühtivad (*söötma/sööma*).

Viitadena on kasutusel:

jätma1 **causes** jääma1, olema5, püsima2
jääma1, olema5, püsima2 **is_caused_by** jätma

Kausatiivsuse leksikaalne markeerimine on eesti verbide seas suhteliselt regulaarne *ta*-liite abil, mis on eesti keele kõige sagedasem verbiliide (Kasik 1996). *ta*-liide võib esineda ka faktitiivses tähenduses, kuid vähemalt selle liite esinemine verbitüves annab põhjust otsida kausatiivsussuhet markeerimata tüvega:

liigutama2 **causes** liikuma3
liikuma3 **is_caused_by** liigutama2

6.3.7. Osasündmuse suhe

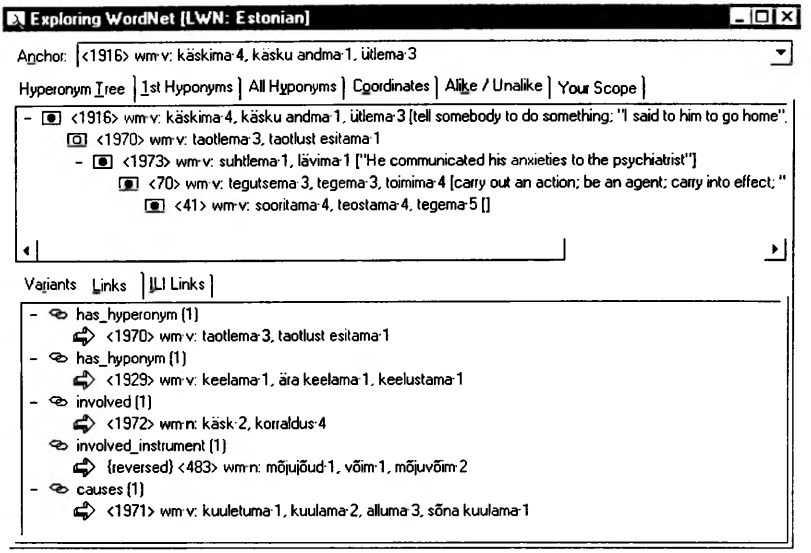
Mitmed verbiga kirjeldatavad situatsioonid jagunevad omakorda
- väiksemateks situatsioonideks (nt *pesu pesema* ja *leotama* ning *loputama*);
- ajaliselt kaasnevateks situatsioonideks (nt *magama* ja *norskama*).

Selline semantiline suhe pole üksühene, see tähendab, et kui suhe kehtib ühes suunas (nt *norskama* on *magama* osasündmus), ei kehti ta tingimata teises suunas (nt *magama* osasündmus pole tingimata *norskama*).

Osasündmuse märkimiseks on kasutusel viidad:

otsustama3, hindama2, hinnangut andma1 **has_subevent** arvama2
õppima2, tudeerima1, koolitükke tegma1 **is_subevent_of** õppima1

Semantiliste suhete tegelikku väljanägemist sisestusliideses Polaris esitab joonis 3 järgmisel leheküljel. Tabel 3 annab kokkuvõtte eesti keele tesauruses esindatud semantilistest suhetest.



Joonis 3. Semantiliste suhete väljanägemine sisestusliideses Polaris

Tabel 3. Eesti keele tesauruses esindatud semantilised suhted seisuga 15. mai 2000

Semantilise suhte nimi	Nimisõna-mõistetes	Verbi-mõistetes	Kokku
NEAR_SYNONYM	22	60	82
XPOS_NEAR_SYNONYM	119	59	178
HAS_HYPERONYM	6353	2479	8832
HAS_XPOS_HYPERONYM	0	6	6
HAS_HYPONYM	6353	2479	8832
HAS_XPOS_HYPONYM	6	0	6
ANTONYM	79	20	99
NEAR_ANTONYM	44	38	82
XPOS_NEAR_ANTONYM	2	2	4
HAS_HOLONYM	17	0	17
HAS_HOLO_LOCATION	4	0	4
HAS_HOLO_MADEOF	6	0	6
HAS_HOLO_MEMBER	9	0	9
HAS_HOLO_PART	57	0	57
HAS_HOLO_PORTION	2	0	2
HAS_MERONYM	17	0	17
HAS_MERO_LOCATION	4	0	4
HAS_MERO_MADEOF	6	0	6
HAS_MERO_MEMBER	9	0	9
HAS_MERO_PART	57	0	57

Semantilise suhte nimi	Nimisõna-mõistetes	Verbi-mõistetes	Kokku
HAS_MERO_PORTION	2	0	2
INVOLVED	7	50	57
INVOLVED_AGENT	2	6	8
INVOLVED_INSTRUMENT	0	21	21
INVOLVED_LOCATION	1	1	2
INVOLVED_PATIENT	1	3	4
INVOLVED_TARGET_DIRECTION	1	3	4
ROLE	57	0	57
ROLE_AGENT	8	0	8
ROLE_INSTRUMENT	21	0	21
ROLE_LOCATION	2	0	2
ROLE_PATIENT	4	0	4
ROLE_TARGET_DIRECTION	4	0	4
CAUSES	6	55	61
IS_CAUSED_BY	27	34	61
IS_SUBEVENT_OF	3	15	18
HAS_SUBEVENT	4	14	18

6.4. Seostamine InterLingual Indexiga

Nagu eespool juba mainitud, on erinevate keelte wordnetid omavahel seotud keeltevahelise indeksi ehk *InterLingual Index*iga. Ideaaljuhul peaksid ILI kirjed mõisteid eristama nii täpselt kui vähegi võimalik (Peters jt 1998). Kui kahest keelest, mis wordnetis on, ühes ei tehta kahe mõiste vahel vahet, aga teises tehakse (nt on itaalia keeles üks sõna nii sõrmede kui varvaste kohta), siis koostatakse ILI kirjed selle keele järgi, kus tehakse peenemat vahet mõistete vahel ja lisatakse ka see mõiste, mis hõlmab teises keeles molemat tähendust. Praktikas on ILI kirjed koostatud WN alusel. Neid on hiljem mõned korrad täiendatud ja parandatud, kuid põhinevad nad siiski suuresti inglise keele mõistelisel jaotusel. Kuna nii ILI kirje moodustavad sõnad kui ka seletus (*gloss*) on ingliskeelsed, siis taandub ILI vastete leidmine suurel määral ingliskeelsete vastete leidmisele.

Olulisemad suhted ILI ja mingi keele wordneti vahel on:

1) *eq_synonym* – täpne vaste;
2) *eq_near_synonym* – ligikaudne vaste;
3) *eq_has_hyperonym* – sõna tähendus on spetsiifilisem kui mingi ILI kirje;
4) *eq_has_hyponym* – sõna tähendus on üldisem kui ILI kirje.

Suhete seadmisel on selgelt eelistatuim *eq_synonym* suhe. Kui seda ei saa kasutada, siis proovitakse *eq_has_hyperonym* või *eq_has_hyponym* suhet rakendada, ja kui needki hõlmavad mõistet ebapiisavalt, siis *eq_near_synonym.* Peab tunnistama, et oleme viimatimainitud suhteliiki kasutanud ka juhtumitel, kus inglise wordneti vastava sünohulgas tähendusväli ei kattu eesti suhtestatava sünohulga tähendusväljaga või kui ingliskeelse sünohulga seletus ei sobi eestikeelse vastega. Korrektne on kasutada *eq_near_synonym* suhet siiski vaid juhul, kui sünohulgale on seatud vastavusse rohkem kui üks ILI kirje, muidugi kui pole tegemist hüpo- või hüperonüümiasuhtega.

Mõisteid, millele ei saa anda ühegi ILI kirjega *eq_synonym* suhet, võib jagada laias laastus kahte kategooriasse.

1. Eesti keeles on palju liitsõnu nagu saksa keeleski. Võtkem kasvõi markantse näitena seesama *liitsõna,* mis eesti keeles kirjutatakse kokku, ingliskeelne vaste *compound word* aga lahku. Vähe sellest, ka sellist mõistet nagu *liitsõna* pole ILIde hulgas olemas. *Liitsõna* on suhtestatud järgmisel viisil:

   liitsõna ***eq_has_hyperonym*** *word (noun)* – (a unit of language that native speakers can identify; "words are the blocks from which sentences are made"; "he hardly said ten words all morning")

   liitsõna ***eq_be_in_state*** *compound (adj)* – (consisting of two or more substances or ingredients or elements or parts; "soap is a compound substance"; "housetop is a compound word"; "a blackberry is a compound fruid")

2. Erinevused kultuurilises taustas. Inglise keeles pole leksikaliseerunud paljud mõisted, mis meie jaoks on igapäevased. Olime hädas selliste sõnadega nagu *viisaastakuplaan, lõõtspill, seljanka, hingedeaeg.* Ka sõnadele *kali* ja *mõdu* oli raske vasteid leida, samas kui *õlle* liike oli ILIde hulgas segadusseajavalt palju.

Kultuurilise konteksti, võibolla isegi harjumuspärase elukeskkonna erinevuse süüks saab ajada paljud erinevused mõistete piirides ja nende leksikaliseerumises. Nii näiteks käituvad eesti sõna *linn* ja hollandi sõna *stad* üsna ühtmoodi, tähistades kõike mõnusast väikelinnast (*town, burg*) kuni mäsleva ja pulbitseva metropolini (*city, metropolis*). Samas pole inglise keeles sellist mõistet nagu *asula,* mis hõlmaks nii *linna, küla* kui *alevit. Settlement1, village1* ja *small*

town1 on ühes sünohulgas, millele vastab eesti *alev1*; *küla1* vasteks on *village2, hamlet1*. Seega võib meie mõistes *alevi* kohta öelda inglise keeles nii *village* kui *town*.

Verbide puhul kerkib ka kausatiivsuse probleem. Inglise *to move* vasteteks on nii *liikuma* kui *liigutama*. Viimane neist on kausatiivne, esimene mitte. Eesti keeles on need erinevused süstemaatiliselt leksikaliseerunud, sama ei saa öelda teiste keelte kohta. Leidub juhtumeid, kus ILI kirjetes on kausatiivsetel ja mittekausatiivsetel verbidel vahet tehtud, kuid see pole kõigi verbide puhul nii. Nt verb *to spend (kulutama)* on inglise keeles leksikaliseerunud ainult kausatiivses tähenduses, *kuluma* tõlkevasteks oleks *be spent*, aga eesti keeles ei ole sõna *kuluma* passiivis. Sellised juhtumid on lahendatud suhete *eq_causes* (24 korral) ja *eq_is_caused_by* (tervelt 80 korral) abil, vastavalt sellele, milline pool, mittekausatiivne või kausatiivne, oli äratuntav ingliskeelses seletuses.

Vaadates sünohulki kronoloogilisest aspektist – nad on nummerdatud kasvavalt – võib täheldada, et esimese tuhandekonna sõna kattuvus ILI kirjetega on parem kui järgmisel. See on tingitud erinevast metoodikast, kuna esimeste sünohulkade puhul olid ingliskeelsed baasmõisted kõigile projektiosalistele ette antud, neile leidsime lihtsalt eestikeelsed vasted.

7. Kasu keeletehnoloogiale ja keeleteadusele

Sarnaste leksikaalsemantiliste ressursside olemasolu paljudes keeltes võib viia mitmete heade tulemusteni.

Automaatsed tõlkesõnastikud on ainult üks neist. Erinevate keelte wordnet'e saab ka leksikaal-semantilisest seisukohast kõrvutada ja võrrelda mitmesuguste semantiliste väljade struktuuri. EWN projektis paluti kõigil osalejatel peensusteni välja arendada näiteks muusikariistade, mõõteriistade ja emotsioonide hierarhiad, mida võrreldi ILI kaudu omavahel leksikaliseerituse, hüponüümiatasandite ja paljude muude näitajate poolest. Ühtse sisestusliidese olemasolu võimaldab ka ainult ühes keeles mingeid suuremaid semantilisi välju detailsemalt kujutada ja uurida, selline töö on eesti keele kohta olemas näiteks direktiivsete verbide alal (Orav 1998).

Et mingil keeletehnoloogia rakendusel oleks võimalik teksti "mõista" olgu siis tegemist infootsisüsteemi või masintõlkega, on hädavajalik, et tekstis olevate sõnade tähendused oleksid üheselt

määratud – ehk teisiti öeldes, tekst peab olema **semantiliselt ühestatud**.

Automaatse semantilise ühestamise mingi meetodi väljatöötamine koosneb tüüpiliselt kahest allülesandest. Esiteks tuleb vastavas tekstis väljavalitud sõnad käsitsi ühestada ja teiseks rakendada sama teksti samadele sõnadele vastavat automaatse ühestamise meetodit. Valisime semantiliseks ühestamiseks osa Orwelli teosest "1984" mis on juba varem lemmatiseeritud ja morfoloogiliselt ühestatud projekti MULTEXT-EAST raames.

Katsealuses lõigus on 1000 nimisõna, mis esitavad 544 lemmat (keskmiselt esines üks lemma 1,84 korda), millest 315 lemmat (58%) on eesti keele tesauruses juba olemas ja 229 lemmat (32%) eesti keele tesaurusest puudu. Kas kõiki puuduvaid sõnu on mõistlik tesaursesse lisada, on omaette küsimus, sest paljud neist on konkreetsele teosele ainuomased. Puuduolevate seas on 5 pärisnime, 129 liitsõna ja 26 tuletist.

Esineb 41 sellist tähendust, mida tähistav sõna on küll üldkeele tesauruses olemas, kuid tähendus ise puudu, need oleks vaja küll lisada. Esialgsed tulemused (kuigi saadud ebapiisava tekstihulgaga) näitavad, et kui kasutada wordneti-tüüpi tesaurust ka automaatsel semantilisel ühestamisel, on olemasolev eesti keele tesaurus alles liiga napp. Kas üldse kasutatakse ühe teksti piires sõnu väga erinevas tähenduses? Ilmnes, et ainult 27 lemmat (5% lemmadest) esinesid tekstis 2-3 tähenduses (nt *asi* ja *jõud* olid 3 tähenduses).

Seda, et semantiline ühestamine pole ka inimesele – spetsialistile – lihtne ega enesestmõistetav ülesanne, näitab asjaolu, et kõnealuse teksti esialgsel käsitsiühestamisel kahe tesaurusekoostaja poolt ühestati erinevalt 80 tekstis esinenud lemmat. Erinevused said omavahel läbi arutatud ja ühe variandi peale kokku lepitud, kuigi pole selgeid kriteeriume, millises kontekstis millist tähendust eelistada. Erinevuste seas oli hulgaliselt ka juhtumeid, mil üks ühestajaist arvas, et tekstis esinev tähendus on veel tesaurusest puudu.

Kirjandus

Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, T., Peters, W. 1998. The linguistic design of the EuroWordNet database. – Special Issue on EuroWordNet. Toim. N. Ide, D. Greenstein, P. Vossen. Vol 32, 2–3, 91–115.

Beckwith, R., Fellbaum, C., Gross, D., Miller, G. 1990. WordNet: A lexical database organized on psycholinguistic principles. – Using On-line Resources to Build a Lexicon. Toim. U. Zernik. Hillsdale, NJ: Erlbaum. 211–231.

Bloksma, L., Diez-Orzas, P L., Vossen, P. 1996. User Requirements and Functional Specification of the EuroWordNet project. Deliverable D001, WP1, EuroWordNet, LE2-4003.

Calzolari, N. 1990. Lexical databases and textual corpora: perspectives of integration for a lexical knowledge base. – Using On-line Resources to Build a Lexicon. Toim U. Zernik. Hillsdale, NJ: Erlbaum. 191–208.

Cruse, D.A. 1986. Lexical Semantics. Cambridge University Press.

Diez-Orzas, P L., Forest, P., Louw, M. 1996. High-level Architecture of the EuroWordNet Database. A Novell ConceptNet-based semantic network. Final version 7. EuroWordNet.

EKSS = Eesti kirjakeele seletussõnaraamat. Tallinn: Keele ja Kirjanduse Instituut/Eesti Keele Instituut. I köide, A–J, Tallinn 1988–1991; II köide, K, Tallinn 1992–1993; III köide, L–N, Tallinn 1992–1994; IV köide, O–rappevili, Tallinn 1994–1996; V köide, 1. ja 2. vihik, rappima–sentimeetririhm, Tallinn 1997–1998.

EuroWordNet-2: Extending EuroWordNet with Other Languages. Telematics Application Programme, Project LE-8328. 1997.

Hallik, T. 1990. Üks katse automatiseerida mõisterühmade moodustamist. – Arvutuslingvistika sektori aastaraamat 1988. Toim J. Ross. Tallinn: ETA Keele ja Kirjanduse Instituut. 56–66.

Kasik, R. 1996. Eesti keele sõnatuletus. Tartu: Tartu Ülikooli Kirjastus.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. 1990. Introduction to WordNet: an on-line lexical database. – International Journal of Lexicography 3, 235–312.

Miller, G., Fellbaum, C. 1991. Semantic networks of English. – Cognition 41, 197–229.

Orav, H. 1998. Eesti keele direktiivverbide semantilise välja struktuur tesaurusena. Magistritöö üldkeeleteaduse alal. Käsikiri. Tartu Ülikool.

Peters, W., Vossen, P., Diez-Orzas, P., Adriaens, G. 1998. Cross-linguistic alignment of Wordnets with an inter-lingual-index. – Computers and the Humanities 32 (2–3), 221–251.

Saagpakk, P. 1992. Eesti–Inglise sõnaraamat. Estonian–English Dictionary. Tallinn: Koolibri.

Vider, K., Orav, H. 1998. Sõna tasandilt mõiste ruumi. – Keel ja Kirjandus 1, 57–64.

Vider, K. 1997 Some problems in Estonian WordNet. – Papers of the Second Swiss–Estonian Student Workshop on Computational and Theoretical Linguistics. Zurich. Electronic publication http://www.cl.ut.ee/ee/yllitised/

Vossen, P., Bloksma, L., Peters, C., Alonge, A., Roventini, A., Marinai, E., Castellon, I., Marti, T., Rigau, G. 1998. Compatibility in interpretation of relations in EuroWordNet. – Computers and the Humanities, 32 (2–3), 153–184.

Vossen, P 1998. Introduction to EuroWordNet. – Computers and the Humanities, 32 (2–3), 73–89.

Õim, A. 1991. Sünonüümisõnastik. Tallinn.

Õim, A. 1995. Antonüümisõnastik. Tallinn.

Õim, H. 1997 Eesti keele mentaalse maailmapildi allikaid ja piirjooni. – Pühendusteos Huno Rätsepale. Tartu Ülikooli eesti keele õppetooli toimetised 7. Toim M. Erelt, M. Sedrik, E. Uuspõld. Tartu. 255–268.

KeeleWeb – http://ee.www.ee/

WordNet – http://www.cogsci.princeton.edu/~wn/w3wn.html

EuroWordNet – http://www.hum.uva.nl/~ewn/

Adjektiivid kui semantiline probleem: wordnet-tüüpi tesauruste koostamise kogemused

Heili Orav
Tartu Ülikool

1. Sissejuhatus

Kõigis keeltes on mõni meetod nimisõnade tähenduste täpsustamiseks või modifitseerimiseks, ehkki nende süntaktiline vorm on erinev.

Eesti, aga ka mitmetes teistes keeltes võib näidata nimisõna omadusi väga mitmekesisel viisil: adjektiividega (nt *suur, soe*), verbi kesksõnavormidega, mis käituvad nagu adjektiivid (nt *kriuksuv uks*), liitsõnadega (nt *tugitool*), pre- ja postpositsioonidega (nt *tool akna juures*) ja nimisõnafraasidega (nt *minu isa kabinet*). Samuti võib kogu osalause nimisõna laiendada (nt *diivan, mille sa ostsid oksjonilt*). Selle artikli tähelepanu on adjektiividel.

Adjektiivide alaseid kirjutisi keeleteaduses on võrdlemisi rohkesti. Semantilisest aspektist vaadelduna nõuab see sõnaliik erilist tähelepanu. Eesti keele adjektiivide käsitlemisel on pööratud senini tähelepanu omadussõnade morfoloogilisele ja süntaktilisele aspektile. Huvi keskmes on olnud komparatiivi moodustamine, adjektiividele orienteeritud lausemallid jms. Paraku ei ole ülevaadet eesti keele adjektiivide semantikast. Käesolevas töös püütakse esitada adjektiivide kui sõnaliigi mitme semantilise eripärasuste tüüpe, mis kerkivad esile praktilise töö – tesauruse koostamise – käigus.

Kõigi sõnaliikide semantilistes eripärasustes ongi jõutud mingile selgusele tegelikult alles suuremate semantiliste andmebaaside – tesauruste – koostamisel, millest esimene ja tuntuim on WordNet. Kuna WordNeti metoodika tundus kõigile huvipakkuv, hakkasid erinevad keeled looma omi semantilisi andmebaase, millede jaoks võeti eeskuju WordNetist. Nimisõnade ja verbide puhul võeti üle sama lähenemine – sünonüümihulgad jagati hierarhiatesse, lisati erinevate sõnaliikide vaheline osaluse, põhjuslikkuse vm. suhe. Adjektiivide kohta on aga erinevatel keeltel erinevad lähenemised. Adjektiivide kui sõnaliigi leksikaalne positsioon on unikaalne ja erineb teistest süntaktilistest kategooriatest, nagu nimisõna ja verbi

vormid. Nendest eripärasustest on tehtud mitmeid uurimusi, mis on jäänud aga kahjuks ainult teoreetiliseks.

1996. aastal alustati Euroopa Komisjoni projektiga EuroWordNet, mille eesmärgiks oli luua WordNeti eeskujul mitmekeelne leksikaal-semantiline andmebaas, milles erinevate keelte wordnetid on ühendatud keeltevahelise indeksi kaudu. Eestist liitus selle projektiga 1998.a. TÜ arvutuslingvistika uurimisrühm. Vastavalt projekti ülesannetele on kaheksa osavõtnud maad koostanud oma keele andmebaasi nimisõnadest ja verbidest – adjektiivid jäeti sellest projektist üldse välja. Põhjuseks see, et adjektiivide käsitlemisel pole sellist üldaktsepteeritavat lahendit nagu teistel sõnaliikidel. Adjektiivide käsitlemisele EuroWordNeti põhimõtete järgi jõuti alles pärast projekti lõppu.

Selles artiklis tutvustan kahe suure leksikaalse andmebaasi – WordNeti ja GermaNeti – lähenemist adjektiividele. Eesti üldkeele tesauruse koostamise seisukohalt on see oluline info, kuna eesti wordneti eeskuju on olnud EuroWordNet, milles adjektiivid puuduvad.

2. Adjektiivid WordNetis

WordNet sisaldab 19500 adjektiivi, mis on organiseeritud 10000 sünohulka. (Ülevaate WordNeti kui semantilise andmebaasi olemusest ja teiste sõnaliikide käsitlemisest WordNetis vaata Kadri Videri artiklit käesolevas kogumikus või Orav, Vider 1998).

Princetoni WordNeti autorid eristavad kahte tüüpi adjektiive: kirjeldavad (*descriptive*) ja relatsioonilised (*relational*) adjektiivid.

2.1. Kirjeldavad adjektiivid

Kirjeldavate adjektiivide all mõeldakse omadussõnu, mis väljendavad millegi omadust ning millel on võrdlusastmed. Neil adjektiividel pole ülem–alamsuhteid, mis on nii iseloomulikud nimisõnadele. Adjektiivid moodustavad pigemini abstraktse hüperruumi, mitte hierarhilise puu. WordNetis on kirjeldavad adjektiivid organiseeritud opositsiooni (antonüümia) ja tähenduse sarnasuse (sünonüümia) järgi.

2.1.1. Antonüümia probleeme

Antonüümia on peamine suhe kirjeldavate adjektiivide vahel. G. A. Miller alustas semantilise andmebaasi loomist psühholingvistiliste testide läbiviimisega, kus inimene pidi ütlema sõnaga esimesena assotsieeruva sõna. Tuntud adjektiivide puhul ütlesid täiskasvanud alati antonüümi (Miller jt 1990).

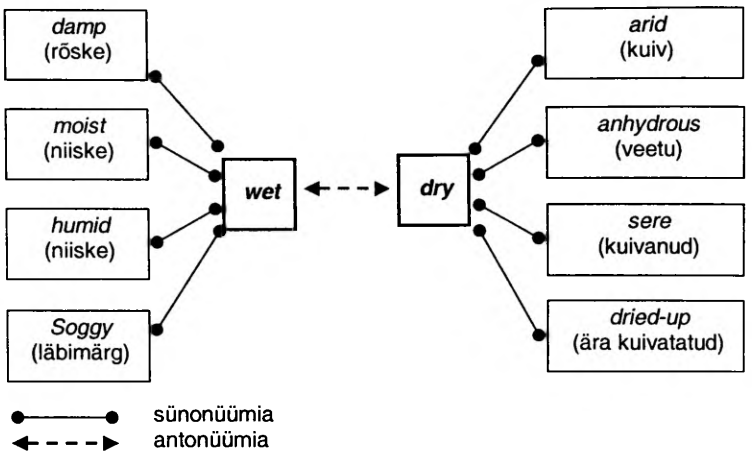
Antonüümia tähtsus on arusaadav, sest adjektiivide tüüpiline ja olulisim funktsioon on väljendada mingi tunnuse väärtusi ja pea kõik atribuudid on mitmepolaarsed. Antonüümsed adjektiivid väljendavad atribuudi vastanduvaid väärtusi.

Sellest teemast ilmneb kaks lähedalt seotud küsimust:

1. Kui kahel adjektiivil on väga sarnane tähendus, siis miks pole neil sama antonüümi (nt. *heavy ja weighty* antonüümid on *light ja weightless*)?
2. Kui antonüümia on nii tähtis, siis miks pole paljudel kirjeldavatel adjektiividel antonüümi? On seal mõni teine semantiline suhe?

Esimene küsimus põhjustas mitmeid probleeme WordNetis, mille eesmärk oli algselt näidata semantilisi suhteid leksikaalsete mõistete (sünohulkade) vahel. Adjektiivide puhul ei ole täpne rääkida antonüümiast sünohulkade vahel. Nt *heavy, weighty, ponderous → light, weightless, airy*. Inimesed, kes oskavad inglise keelt, teavad, et *heavy/light* on antonüümid ja võib-olla ka *weighty/weightless*, aga nad jäävad pead murdma *heavy/weightless* või *ponderous/airy* üle. Mõisted on vastandatud, aga sõnavormid pole tuttavad ega üldtunnustatud antonüümipaarid.

Gross, Fischer ja Miller (1989) pakkusid välja, et adjektiivide sünohulki võiks vaadelda kui adjektiivi "kobaraid" mis assotsieeruvad atribuudi vastandite kobaraga. Nt kui *ponderous* sünonüüm on *heavy* ja *heavy* antonüüm on *light,* siis *ponderous/light* mõisteline vastandus on vahendatud läbi *heavy*. Gross, Fischer ja Miller eristasid otseseid antonüüme nagu *heavy/light,* mis on mõistelised vastandid ja mis pole leksikaalselt paari pandud. Selles formuleeringus on kõigil kirjeldavatel adjektiividel antonüümid – millel puudub otsene (*direct*) antonüüm, on kaudne (*indirect*) antonüüm, st nad on sünonüümid neile adjektiividele, milledel on otsesed antonüümid. Näide illustratsioonil on antonüümipaari märg/kuiv (*dry/wet)* kohta.



See strateegia on edukas suurema osa inglise adjektiivide puhul, aga mõnede puhul tekitab see WordNetis huvitavaid probleeme. Nt tunnet näitav omadussõna *vihane* (*angry*). Erinevalt enamusest atribuutidest ei tundu see olevat bipolaarne – *viha* atribuut pole gradatsiooniline mittevihasest eriti raevununi/maruvihaseni. Ehkki on palju sõnu sarnased vihase tähendusele: ärritatud, raevunud, maruvihane jms, pole ühelgi neist otsest antonüümi. Lähedasem antonüümipaar on *pleased/displeased*, kuid see kaotab ära vihase olulise tähenduse tunde intensiivsusest lähtudes. WordNetis lahendati see probleem sünohulga *not angry* loomisega, millel on lähisünonüümideks *calm* ja *placid*.

2.1.2. Gradatsioon

Teoreetilis-semantiliselt orienteeritud töödes on sellega palju tegeldud.

Enamus diskussioone antonüümiast on vasturääkivuse/kontradiktoorsuse (*contradictory*) ja vastandlikkuse/kontraarsuse (*contrary*) ümber. See terminoloogia on pärit loogikast. Nt *elus* ja *surnud* on vasturääkivad, sest tõene väide *Kennedy on surnud* eeldab, et on väär, et *Kennedy on elus* ja vastupidi. Vastandlikud mõisted on nt *kuum* ja *külm*. Laused *Vesi on kuum* ja *Vesi on külm*, ei saa mõlemad olla tõesed, ehkki mõlemad saavad olla väärad kui nt vesi on soe.

Lyons (1977) juhtis tähelepanu sellele, et vastandlike mõistete definitsioon ei piirdu ainult opositsiooniga, vaid seda saab rakendada nii laialt, et ta muutub peaaegu mõttetuks. Nt *Kennedy on puu* ja *Kennedy on koer* mõlemad ei saa olla tõesed, aga mõlemad saavad olla väärad – järelikult peavad *puu* ja *koer* olema vastandid. Lyons väidab, et see on gradeeritavus, mitte tõesus, mis annab parema seletuse neile erinevustele. Gradeeritavad adjektiivid on vastandlikud, mitte vasturääkivad.

Gradatsiooni peaks käsitlema kui semantilist suhet, mis organiseerib adjektiive leksikaalses mälus. Gradatsioon on oma süvaolemuselt võrdlemine, sest mingi objekti omaduse intensiivsusaste on aste vaid teiste objektide vastava omaduse intensiivsusega võrreldes. Mõne atribuudi gradatsioon võib väljenduda järjestatud adjektiividega, mis kõik viitavad ühele ja samale nimisõnale WordNetis.

Tabel illustreerib leksikaalset gradatsiooni suuruse, vanuse ja soojuse kohta.

Size	Age	Warmth
astronomical	ancient	torrid
huge	old	hot
large	middle-aged	warm
standard	mature	tepid
small	adolescent	cool
tiny	young	cold
infinitesimal	infantile	frigid

Kõige raskem on leida nimetusi neutraalse keskmise kohta – äärmused on ulatuslikumalt leksikaliseerunud. See tabel on rohkem erand kui reegel ja näitab, et inglise keeles on gradatsioon üsna vähe leksikaliseerunud. Selleks kasutatakse rohkem adverbe – *very, quite, pretty, more, most* jne. Selliseid suhteid sünohulkade vahel võiks olla D. Grossi jt. (Gross, Fellbaum, Miller 1993) hinnangutel maksimaalselt 2% enam kui 2500 adjektiivi kohta. Seega kontseptuaalselt oluline ja teoreetikuid huvitav gradatsiooni suhe ei mängi adjektiivide organiseerimisel keskset rolli ja pole seetõttu WordNeti sisse võetud.

Eesti keele adjektiivide gradatsiooniga on tegelenud M. Erelt (1986). Tema huviks on aga peamiselt olnud adjektiivide kompa-

ratsioonisüsteemi süntaktiline iseärasus, seega pole teada, kui palju on gradatsioon leksikaliseerunud eesti keeles.

2.1.3. Polüseemilisus

Keeleteaduses oli Edward Sapir (1944) esimene lingvist, kes osutas, et paljud adjektiivid omandavad erineva tähenduse, kui nad laiendavad erinevat nimisõna. Nt *pikk* on erineva tähendusega maja, puu või inimese kohta kasutatult.

Adjektiivid on valivad nimisõnade suhtes, mille tähendusi nad määratlevad. Osad adjektiivid on üldkasutatavad peaaegu igas kontekstis (nagu hinnangulised adjektiivid *hea/halb; soovitud/soovimatu*) koos iga nimisõnaga. Samuti on lai kasutusala tegevust või võimsust märkivatel omadussõnadel (nt *kiire/aeglane; tugev/nõrk*). Mõned adjektiivid aga on väga tugevasti seotud nimisõnaga, mida nad laiendavad (nt *niidetud/niitmata*).

WordNetis on võetud seisukoht, et mitmetähenduslike adjektiivide mitmetähenduslikkust aitavad vähendada antonüümid (nt *värske : must (särk); värske : vana (leib); värske : hapu (piim)*).

2.2. Värviadjektiivid

Need adjektiivid on erakordsed mitmel viisil ja neid on palju uuritud. Aga nad moodustavad keeles ka väga spetsiifilise rühma ja nende uurimisel saadud tulemusi on raske üldistada teistele adjektiivide rühmadele.

Inglise keeles on nad käsitletavad nii kirjeldavate adjektiivide kui nimisõnadena. Aga see, mis neid eristab, on antonüümia – neil pole otseseid ja mitteotseseid antonüüme. Ainult üht värvuse tunnust saab kirjeldada otsese antonüümi kaudu: heledus (*lightness*), mille äärmised vastandused on *hele/tume* (nt *helepunane : tumepunane*). On ka palju teisi vastandusi, mida võiks vaadelda kui opositsioone (nt *sinine : punane*). Selliseid lekseeme on nt Katz aastail 1964 ja 1966 käsitlenud kui antonüüme, kuid see on ebatavaliselt lai interpreteering terminile "antonüüm"

WordNetis vastandus *värviline/värvitu* hõlmab inglise keele värvide nimetusi, st värvid (nt *sinine, kollane* jt.) on kodeeritud kui *värvilise* sünonüümid.

Eesti värviadjektiive on uurinud U. Sutrop (1995, 1996). Tema tähelepanu on värvisõnavaral ja eelkõige selle ajaloolisel arengul.

2.3. Relatsioonilised adjektiivid

Need adjektiivid on leidnud lähemat uurimist näiteks Levi (1978) poolt ja tähendavad 'on seotud/lubatavad või assotsieeruvad' mõnede nimisõnadega ja mängivad nendega sarnast rolli.

Nt *vennalik – vennalikud kaksikud* on seotud vennaga; *hamba- – hambahügieen* seotud hammastega jne.

Mõned nimisõnad on aluseks mitmele homonüümsele adjektiivile: inglise keeles nt *musical instrument ja musical child.*

Neil adjektiividel pole otseseid antonüüme ja neist ei saa moodustada rühmi nagu kirjeldavatest adjektiividest. Nende vastandus võib tulla nimisõnast, mida nad kirjeldavad. Nt *kriminaalne* vastandub *tsiviilõiguslikule seaduste* kontekstis.

WordNetis on 1700 relatsioonilist adjektiivide sünohulka, mis sisaldavad rohkem kui 3000 lekseemi. Iga sünohulk viitab vastavale nimisõnale. Nt sünohulk *stellar, astral, sidereal* viitab sõnale *star.*

3. Adjektiivid GermaNetis

3.1. GermaNet

GermaNet on elektrooniline leksikograafiliste viidetega andmebaas saksakeelsete mõistete jaoks. Ta on oma olemuselt kokkulangev Princetoni WordNetiga, kuigi sisaldab põhimõttelisi ja organisatoorseid modifikatsioone nii leksikaalsete kui kontseptuaalsete suhete tasemel ja sisaldab muuhulgas just adjektiivide osas mitmeid huvipakkuvaid ideid ja lahendusi WordNetiga võrreldes. Uus on GermaNetis ka nt lähenemine regulaarsele polüseemiale, nn kunstlikele mõistetele (selle eristamiseks kasutatakse spetsiifilist märgendit) ja verbi partitsiibi vormidele (Hamp, Feldweg 1997).

GermaNetis on samuti nagu WordNetis andmebaasi jaotus nelja sõnaliigi vahel: nimisõnad, verbid, adjektiivid ja adverbid. Iga sõnaliigi semantiline ruum on GermaNetis jagatud 15 semantilisse välja. Sellise jaotuse eesmärk on peamiselt organisatoorne, see võimaldab töö tesauruse kallal jagada osadeks, et leksikograafil oleks hõlpsam vastavat andmefaili toimetada.

Sakslased eristavad kahte põhilist suhete tüüpi: leksikaalsed ja kontseptuaalsed suhted. Leksikaalsed suhted – sünonüümia ja antonüümia – on kahesuunalised (*bidirectional*) suhted, mis kehtivad kõigi sõnaliikide kohta. Ülejäänud suhted (hüponüümia, mero-

nüümia jm) on kontseptuaalsed suhted ja erinevad vastavalt sõnaliigile.

GermaNetis rõhutatakse **erinevate sõnaliikide vaheliste suhete** tähtsust. Mõnda suhet kasutatakse märksa sagedamalt kui WordNetis, mõne suhte kasutust on laiendatud ja on loodud ka uusi sõnaliikide vahelisi suhteid. Nt põhjuslikkuse (*cause*) suhe, mida WordNetis kasutatakse ainult verbide vahelisena, on GermaNetis võetud kasutusele ka verbide ja adjektiivide vahelisena (nt avama (*öffnen*) põhjustab avatud (*offen*) olemise).

Väga suurt tähelepanu pööratakse GermaNetis **regulaarsele polüseemiale**. Selleks eristatakse GermaNetis uut kahesuunalist suhet. Nii on nt *pank* suhestatud samaaegselt nii *asutuse* kui ka *ehitisega* või *siga* on suhestatud nii *looma* kui ka *toiduga*. Sama suhet kasutatakse ka nt SIMPLE projektis, mis on PAROLE projekti jätk ja mille üheks eesmärgiks on hinnata erinevate Euroopa keelte regulaarset polüseemiat ja seda leksikaalsetes andmebaasides ühtse meetodi järgi vähendada (Bel jt 2000).

Adjektiivid on GermaNetis leidnud täiesti uue lähenemise, milleks on adjektiivide **taksonoomia.**

3.2. Adjektiivide semantilised väljad

Inglise WordNetis pole adjektiivid klassifitseeritud semantilistesse väljadesse nagu nimisõna ja verbi, pigem on koondatud kõik adjektiivid ühte tohutusse faili. GermaNet aga eristab adjektiivide vahelisi erinevaid semantilisi klasse. GermaNetis on 1613 adjektiivide sünohulka, mis on jaotatud 14 semantilisse klassi ja lisaks on 711 sünohulgast koosnev **pertonüümide** (ingl *pertainym;* sks *pertonym*) klass. Sellise semantiliste klasside jaotuse aluseks võtsid GermaNeti tegijad Hundsnurscheri ja Spletti (1982) uurimuse. Enne Hundsnurscheri ja Spletti kasuks otsustamist uuriti ka teiste lingvistide uurimusi (nt Rachidi, Dixon, Lee jt) Niisiis on selline jaotus enam lähtuv teatud teoreetilisest mudelist. Mõned muudatused on tehtud mõne semantilise klassi allkategooriates, lisaks on, nagu juba öeldud, moodustatud eriline nn. pertonüümide klass.

Järgnevalt adjektiivide semantilistest klassidest GermaNetis. Lähemalt kommenteerin ainult mõnda klassi.

• Tajuadjektiivid (*adj.Perzeption*)

Tajuadjektiivide klass sisaldab adjektiive, mis kirjeldavad kõike seda, mida inimene oma viie meelega tajub. Vastavalt sellele, mida me näeme (valgus, värvus, pealispind), kuuleme (heli), maitseme, haistame ja puudutame (pealispind), saab neid adjektiive jagada järgnevasse alljaotusesse:

1) heledus (*Helligkeit*);
2) värvus (*Farbe*);
3) heli (*Geräusch*);
4) maitse (*Geschmack*);
5) lõhn (*Geruch*);
6) pealispind (*Oberfläche*).

Adjektiivid, mis kirjeldavad materjali pealispinda, osutavad nii seda, mida me saame katsuda (nt *pehme, sile*) kui seda, mida me näeme (nt *läikiv, matt*).

• Ruumiliste omadustega seotud adjektiivid (*adj.Ort*)
• Aja omadustega seotud adjektiivid (*adj.Zeit*)
• Liikumisega seotud adjektiivid (*adj.Bewegung*)
• Materjaliga seotud adjektiivid (*adj.Substanz).*
• Ilmaga seotud adjektiivid (*adj.natPhaenomen*)
• Kehaga seotud adjektiivid (*adj.Koerper*)
• Meeleoluga seotud adjektiivid (*adj.Gefuehl*) – selles klassis eristatakse tundeid (*Empfindung/Gefühl*) (nt *rõõmus, õnnelik, vihane* jt) ja stiimuleid (*Reiz*) (nt *mõnus, tüütu* jt).
• Vaimuga seotud adjektiivid (*adj.Geist*)
• Käitumisega seotud adjektiivid (*adj.Verhalten*)
• Sotsiaalseid suhteid iseloomustavad adjektiivid (*adj.Gesellschaft*)
• Hulgaga seotud adjektiivid (*adj.Menge*)
• Suhetega seotud adjektiivid (*adj.Relation*)
• Üldised adjektiivid (*adj.Allgemein*)

Need on "üldise" tähendusega adjektiivid, millel on väga lai kasutus. Nad võivad olla ülemmõisteks paljudele teistele adjektiivide klassidele. Mitme semantilise klassi adjektiivid võivad olla seotud nt hea või halvaga (*gut v. schlecht*). Nt *halb* (*schlecht*) on suhestatud isiku spetsiifilise adjektiiviga *õel/paha* (*böse*), mille hüponüümideks on *alatu, madal, salalik* jms.

• Pertonüümid (*adj.Pertonym*)
Pertonüümide all mõeldakse GermaNetis nimisõnast tuletatud omadussõnu, nagu *finantsiline, intellektuaalne* jms. Paljusid neist ei saa jaotada eelpool kirjeldatud semantilistesse klassidesse, seega on need eristatud omaette klassi. Need tuletatud adjektiivid, mis siiski sobivad mõnda semantilisse klassi, on GermaNetis ka vastavasse faili lisatud. Nt *intellektuaalne* kuulub vaimuga seotud adjektiivide klassi. Iga pertonüüm on seostatud nimisõna või verbiga leksikaalse suhte '**derived from**' (tuletatud) kaudu, nt *finantsiline* derived_from *finantsid.*

Adjektiivid, mis sisaldavad produktiivseid tuletuslikke elemente, on GermaNetti sisestatud ainult siis, kui neil on korpustes või muudes sagedusloendites kõrge esinemissagedus.

Pertonüümide semantilisse klassi kuuluvad ka tuletatud adjektiivid, mis märgistavad millegi eitust või puudumist ja mida GermaNetis nimetatakse **privatiivideks** (sks *Privativa*). Neid võib defineerida järgnevalt "olema mitte x / olema ilma x" nt *plekitu, lõhnatu = lõhnata (gaas).*

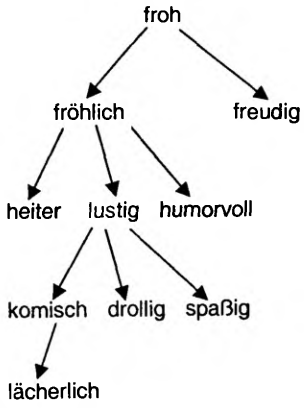
Sarnaselt pertonüümidele on nad selles klassis ainult juhul, kui nad ei kuulu mujale semantilisse klassi.

3.3. Adjektiivide semantilised suhted GermaNetis

GermaNeti tegijad on loobunud adjektiivide jagamisest kirjeldavateks ja relatsioonilisteks adjektiivideks, sest nende eristus pole G. A. Millerit (1993) tsiteerides väga selge: "Otsus selle kohta, millises failis mingi adjektiiv peaks olema, on lõppude lõpuks pragmaatiline" Ka on põhjuseks see, et paljud ingliskeelsed relatsioonilised adjektiivid on saksa keeles tihti realiseerunud liitomadussõnade näol. Sama esineb ka eesti keeles (nt *muusikainstrument*).

Samuti on sakslased loobunud WordNeti nn "sünohulkade kobarate" moodustamisest, sest kaudsete antonüümide kontseptsioon tundub üsna ebaselge. GermaNetis pole seetõttu tekitatud "kunstlikke" antonüüme nagu nt WordNetis *pregnant → nonpregnant!* (*rase → *mitterase*). Selle asemel on püütud GermaNetis sarnased sünohulgad struktureerida hierarhiliselt. Hüponüümia suhet on kasutatud kõikjal, kus see vähegi võimalik.

Näide: "froh" (rõõmus)



Seda hierarhiat vaadates tundub, et see pole päris korrektne hüponüümia suhe. Hüponüümiasuhte kindlakstegemisel saab kasutada Cruse (1986) leksikaal-semantiliste seoste kontrollimiseks pakutud freimi:

> X on teatud Y või
> Kui see on X, siis peab ta olema ka Y

Ehk siis

> Kui see on koomiline, siis peab ta olema ka rõõmus.

See tähendab, et selles hierarhias on kokku pandud adjektiivid erinevatest semantilistest klassidest. Üleval pool on tegemist meeleolu/tunnetega seotud adjektiividega, allpool – alates lõbusast (*lustig*) – käitumisega seotud adjektiividega, st nende tähenduse sisu muutub. Lõbusa, rõõmsa, lustliku (*lustig*) hüponüümid koomiline (*komisch*), naljakas (*spaßig*) ning naeruväärt (*lächerlich*) jt on kasutatavad juba mingi objekti, kellegi käitumise või olukorra kohta. On ilmne, et toodud näitega sarnased hierarhiad vajavad korrigeerimist.

GermaNetis on adjektiivide kirjeldamiseks kasutatud järgmisi semantilisi suhteid.

- Antonüümia (nt *schlecht, gut*) – kasutatakse ainult otseseid antonüüme.
- Hüponüümia (nt *toll, gut*) – seda suhet on püütud rakendada kõikide adjektiivide vahel, kus vähegi võimalik.
- 'Vaata ka' (nt *bundesweit, Bundesland*) – see on pärisnimede ja sellest tuletatud adjektiivide vaheline suhe.

- 'Tuletatud' (nt *medizinisch, Medizin*) – see suhe on adjektiivide ja nimisõnade vaheline. Princetoni WordNetis on need relatsioonilised adjektiivid (vt eespool).
- Kesksõna (nt *bedeutend, bedeuten*) – kesksõna on verbi infiniitne vorm, mis väljendab tegevust omaduse või seisundina. Seetõttu on kesksõna sõnaliigilt lähedane omadussõnadele ja saab esineda täiendina. Tesaurusesse on sisestatud ainult kõige sagedasemalt kasutatavad kesksõnad.

4. Eesti üldkeele tesaurus

Eesti üldkeele tesaurusesse on lisatud praegu (juuni 2000) umbes 200 adjektiivi. Alustatud sai korpuse põhjal tehtud sagedusloendi kõige sagedasematest adjektiividest ja lisatud neile enim assotsieeruvad omadussõnad. Selleks võis olla kas otsene antonüüm või lähisünonüüm.

Kuidas siduda eesti keele adjektiive semantilistesse suhetesse – kas WordNeti eeskujul nn "sünonüümseteks kobarateks" millel on otsesed ning mitteotsesed antonüümid, või püüda jaotada adjektiive hierarhiliselt semantiliste väljade kaupa, nagu on seda tehtud GermaNetis – see uurimine ootab eesti keele tesauruse tegijaid ees.

WordNeti meetodi vastu räägib asjaolu, et nii nagu saksa keeles, realiseeruvad ka eesti keeles relatsioonilised adjektiivid tihti liitomadussõnadena (nt *muusikainstrument, aatompomm*). Samuti oleks otstarbekas vältida kaudsete antonüümide kasutamist, sest väga tihti peaks sel juhul tekitama nn tehisadjektiive. Samas on see oluline suhe, mis vähendab tunduvalt sõnade polüseemilisust. Kui loobuda kaudsest antonüümiast, peab leidma selle asemele uue seda rolli täitva suhte.

Panna adjektiivid GermaNeti eeskujul semantiliste klasside kaupa hierarhilistesse suhetesse on samuti üsna problemaatiline. Juba nimisõnade hierarhiliste suhetega on raskeim küsimus kõige tipmise või tipmiste hüperonüümide leidmine. Samuti ei leia kõige sagedasemate verbide seast ühist ülemmõistet nii transitiivsetele kui intransitiivsetele verbidele. Adjektiividega on lugu veel segasem – pole olemas lekseeme, millede alla koonduks palju omadussõnu.

Eelpool sai mainitud, et WordNeti mõistes relatsioonilised adjektiivid on eesti keeles leksikaliseerunud liitomadussõnadena. Siiski viitavad ka mitmed lihtomadussõnad kindlatele nimisõnadele.

Nt tundub eesti keeles väga vale küsida: "Oli see punane või teistmoodi värviline?" pigem küsime "Oli see punane või mõnda teist värvi?" Samamoodi näitavad erinevad adjektiivid erinevaid kuju, maitse, hääle, vanuse, psühholoogilise seisundi vm omadusi. Lyons (1977) nimetab sellist adjektiivi ja nimisõna vahelist suhet **kvaasiparadigmaatiliseks** (*quasi-paradigmatic*) suhteks. Ta väidab, et kui hierarhiliselt struktureeritud sõnastikus oleks lisaks hüponüümidele ka nn **kvaasihüponüümid** (*quasi-hyponym*), oleks kõigis keeltes kogu sõnavara paigutatud suhteliselt väikese hulga üldise tähendusega lekseemide alla. Omaette probleem on aga see, et kvaasihüponüümia piirid tuleb sel juhul defineerida. WordNetis on see suhe nimetatud **atribuudi** suhteks, nt nimisõna *kaal* on atribuut, mille väärtusi väljendavad adjektiivid *kerge* ja *raske*. EuroWordNetis on nimisõnade ja adjektiivide vaheline suhe ümber nimetatud **be_in_state** ja **state_of.** Nt: {*värv, värvus*} be_in_state {*värviline*} või {*värvitu*}.

5. Kokkuvõte

Eelnenust peaks ilmnema, et adjektiividega seostub semantika jaoks mitmeid spetsiifilisi probleeme.

Mõningaid adjektiivide rühmi (värviadjektiive, maitseadjektiive jt) on küll palju uuritud, aga enamasti mingi üldisema probleemi seisukohalt (nt keele ja mõtlemise seos). Adjektiivide kui sõnaliigi semantika tervikuna on aga praktiliselt käsitlemata.

Siinse artikli eesmärk oli kahene: osutada mõnedele olulisematele adjektiivide tähendusiseärasustele ja teiseks, lähtudes vajadusest lülitada adjektiivid eesti keele semantilisse andmebaasi, käsitleda probleeme, mida need tähenduste iseärasused sellise andmebaasi loomisel tekitavad.

Üheseid lahendusi neile probleemidele artiklis ei pakuta, selleks on vaja eesti keelest lähtuvat põhjalikumat uurimist.

Kirjandus

Bel, N., Busa, F., Calzolari, N., Gola, E., Lenci, A., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., Zampolli, A. 2000. SIMPLE: A general framework for the development of multilingual lexicons. – Second International Conference on Language Resources and Evaluation. Athens.

Cruse, D.A. 1986. Lexical Semantics. Cambridge University Press.

Erelt, M. 1986. Eesti adjektiivisüntaks. Tallinn: Valgus.

Gross, D., Fischer, U., Miller, G. A. 1989. The organization of adjectival meanings. – Journal of Memory and Language.

Gross, D., Fellbaum, C., Miller, K. 1990/1993. Adjectives in WordNet. – International Journal of Lexicography 3 (4), ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Hamp, B., Feldweg, H. 1997. GermaNet – a lexical-semantic net for German. – Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid.

Hundsnurscher, S. 1982. Semantik der Adjektive im Deutschen: Analyse der semantischen Relationen. Westdeutscher Verlag.

Lyons, J. 1977 Semantics 1. Cambridge University Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. 1990. Introduction to WordNet. An on-line lexical database. – International Journal of Lexicography 4.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. 1993. Five Papers on WordNet. Technical Report, Cognitive Science Laboratory, Princeton University. Revised version. ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Orav, H., Vider, K. 1998. Sõna tasandilt mõiste ruumi. – Keel ja Kirjandus 1, 57–64.

Sapir, E. 1944. Grading: A Study in Semantics Philosophy of Science.

Sutrop, U. 1995. Eesti keele põhivärvinimed. – Keel ja Kirjandus 12, 797– 808.

Sutrop, U. 1996. Värvisõnad: ääremärkusi Taani hindamisraamatu Eestimaa lehtede kohta. – Keel ja Kirjandus 4, 225– 229.

Sutrop, U. 1996. Eesti keele värvussõnavara arengu põhijooni. – Keel ja Kirjandus 10, 661–674.

Adjectives in GermaNet. http://www.sfs.nphil.uni-tuebingen.de/lsd/Adj.html

Kasutajaliides info hankimiseks elektroonilisest käsiraamatust: Zürichi ja Tartu ühisprojekt

Neeme Kahusk, Kaarel Kaljurand, Mare Koit, Kadri Vider
Tartu Ülikool

1. Sissejuhatus

Arvutiprogramm, mis suudaks tuvastada loomuliku keele teksti tähendust ja seeläbi vastata loomulikus keeles esitatud küsimustele, oleks kasulik mitmesugustes praktilistes rakendustes, olgu siis tegu info hankimisega ühestainsast mahukast elektroonilisest dokumendist või veebis paiknevast dokumentide massiivist.

Kasutajale on mõistagi kõige mugavam täisautomaatne sisupõhine küsimustele vastamine. Selliste küsimus-vastussüsteemide loomine on aga väga töömahukas, seetõttu eksisteerib neid praegu ainult väga kitsaste ainevaldkondade jaoks ja nad suudavad käsitleda suhteliselt lühikesi dokumente (Herzog, Rollinger 1991).

Suvalisse valdkonda kuuluvatest suurtest tekstihulkadest informatsiooni automaatseks leidmiseks on välja töötatud kaks põhilist lähenemisviisi, mida nimetatakse vastavalt infootsinguks (*information retrieval*) ja info ekstraheerimiseks (*information extraction*) (Mollá Aliod jt 1998). Tüüpilised infootsingu meetodid, mis on realiseeritud näiteks veebi otsingumootorites, võimaldavad kiiresti leida päringule vastavaid dokumente hiigelsuurtest tekstikogumitest. Kui aga dokumendid on mahukad, siis tuleb kasutajal näha palju lisavaeva, et leida tekstist üles asjakohased laused või lõigud. Lisaks sellele on kõigil infootsimeetoditel rida piiranguid, mis teevad nad ebasobivaiks mitmetes tähtsates rakendustes. Esiteks, arvesse võetakse ainult dokumendi sisusõnu ja ignoreeritakse funktsioonisõnu. Seetõttu ei tehta vahet päringutel

> export from USA to Germany,
> export from Germany to USA.

Teiseks, paljudel juhtudel kasutatakse ainult sisusõnade tüvesid, mis leitakse lihtsustatud algoritmi järgi, tegemata täielikku morfoloogilist analüüsi, ja seetõttu mõnikord vigaselt.

Kolmandaks, päringus sisalduvaid sõnu käsitletakse kui järjestamata hulka. Seetõttu ei tehta vahet päringutel

computer design,
design computer.

Neljandaks, ei tehta vahet homonüümidel. Nt päring *banking* annab tulemuseks nii dokumendid, kus käsitletakse lennuki külgkaldega sõitu kui ka dokumendid, kus on juttu pangandusest.

Ka info ekstraheerimise meetodid võimaldavad leida infot suurtest tekstikogumitest ja seejuures suudavad edasi anda olulist infot, kuid eeldusel, et see on eelnevalt defineeritud. Näiteks terrorismiakte kirjeldavatest dokumentidest võib leida andmeid ründaja, ohvri, kasutatud relva jms. kohta. Selline eeldefineerimine on töömahukas ja valdkonnaspetsiifiline, mistõttu info ekstraheerimise süsteem suudab vastata üksnes kindlat tüüpi küsimustele.

Mõistlik kompromiss ühelt poolt täisautomaatse teksti mõistmist eeldava küsimustele vastamise ning teiselt poolt infootsingu ja info ekstraheerimise vahel on meetod, mida nimetatakse vastuse ekstraheerimiseks. Kasutaja esitab päringu, mille alusel leitakse need kohad dokumentides, mis sisaldavad vastuse.

Zürichi ülikoolis on välja töötatud vastuse ekstraheerimise süsteem ExtrAns, mis suudab vastata operatsioonisüsteemi Unix kohta käivatele küsimustele, kasutades elektroonilist käsiraamatut (Hess 1998). Näiteks kui kasutaja küsib

Does cp copy files?

siis väljastatakse talle järgmised tekstilõigud (vt demoversiooni http://www.ifi.unizh.ch/CL/extrans/):

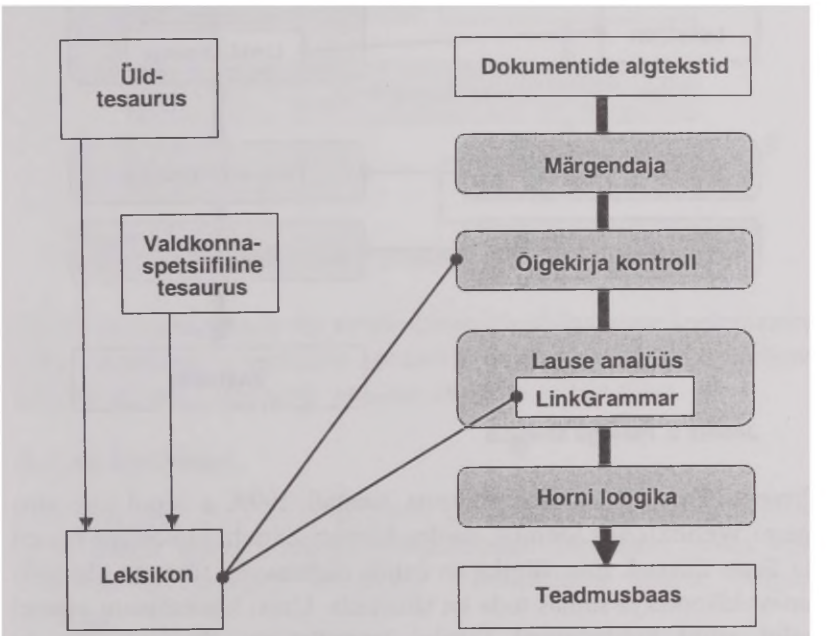
For each source_dir, cp will copy all files and subdirectories
cp – copy files

millest ta saab lugeda vastuse oma küsimusele.

2. Vastuse ekstraheerija

Süsteemi ExtrAns töö toimub kahes etapis. Esimene etapp on ettevalmistav, selle käigus luuakse teadmusbaas Unixi elektroonilise käsiraamatu automaatse analüüsimise tulemusel: kõigepealt analüüsitakse iga lause morfosüntaktiliselt, seejärel lemmatiseeritakse, ühestatakse, lahendatakse anafoorid ja lõpuks leitakse loogiline vorm. Morfosüntaktiline analüüs viiakse läbi sõltuvusorienteeritud süsteemi Link Grammar abil, mille koosseisu kuuluvad sõnavormide leksikon, grammatika ja süntaksianalüsaator (Sleator, Temperley

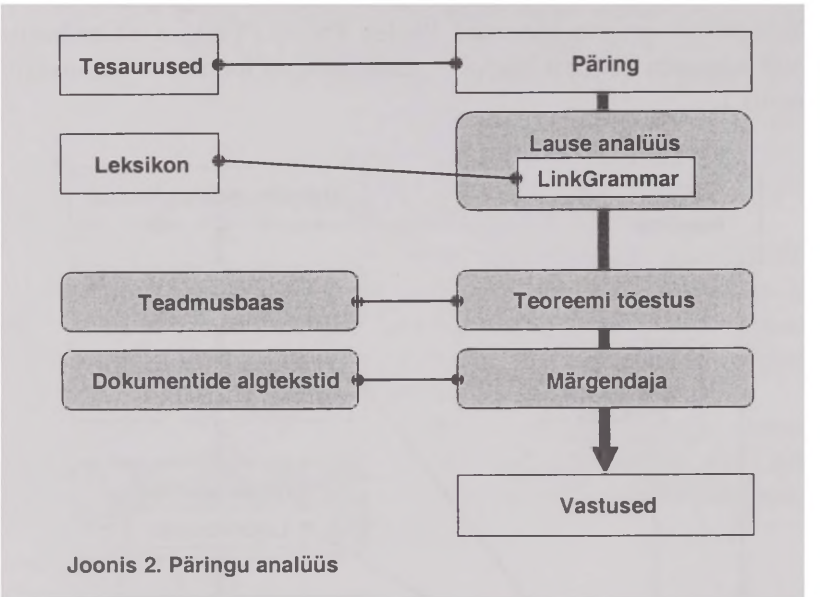
1991). Teadmuse esituse keelena kasutatakse Horni loogikat, kus iga põhiverbi jaoks on sisse toodud fikseeritud kohtade arvuga predikaat. Igale lausele vastab selle lause loogiline vorm – valem, mis sisaldab viidad lause asukohale tekstis ja lauses esinevatele sõnadele. Teadmusbaas on seega Horni valemite hulk, milles viitade abil on säilitatud ka seos elektroonilise dokumendiga (Hess 1997). Süsteem on programmeeritud keeles Prolog. Vastuse ekstraheerimise süsteemi esimest etappi – teadmusbaasi loomist – illustreerib joonis 1.



Joonis 1. Teadmusbaasi loomine

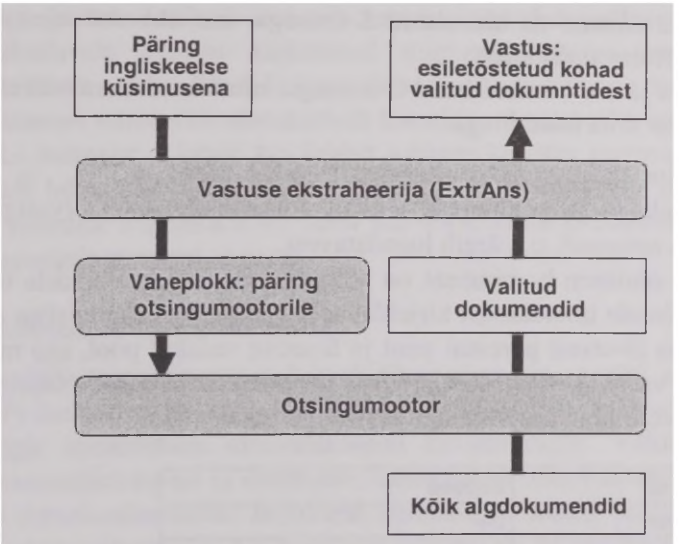
Kasutaja näeb ainult süsteemi ExtrAns töötamise teist etappi, mis funktsioneerib järgmisel viisil. Kasutaja sisestab päringuna inglise-keelse lause. See analüüsitakse morfosüntaktiliselt ja semantiliselt ning saadakse Horni valem. Saadud valemit käsitletakse kui teoreemi, mida püütakse automaatselt tõestada, võttes teadmusbaasi kuuluvad valemid aksioomideks. Tuletusstrateegiana kasutatakse muutuva laiuse ja sügavusega otsingut. Kui "teoreem" õnnestub tõestada, siis tõestuses esinenud loogilistest termidest leitakse viitade alusel laused tekstist, mis kuvatakse ekraanile. Teatavatel

tingimustel kasutatakse otsingu täpsuse ja katvuse tõstmiseks päringu rikastamist sünonüümide ja/või hüponüümidega ning päringus esinevate terminite definitsioone (mõlemad leitakse tesaurusest). Vastuse ekstraheerimise süsteemi teist etappi – päringu analüüsimist ja teoreemi tõestamist – illustreerib joonis 2.



Joonis 2. Päringu analüüs

Projekti ExtrAns edasiarendusena alustati 1999. a lõpul uue süsteemi WebExtrAns loomist, milles kõrvuti Zürichi ülikooliga osaleb ka Tartu ülikool. Eesmärgiks on kanda olemasolev süsteem üle uude ainevaldkonda ja samas teda ka täiustada. Unixi käsiraamatu asemel valiti nüüd kokkuleppel Šveitsi lennufirmaga Swissair lennuki Airbus 320 ingliskeelne tehniline manuaal tekstimahuga umbes 120 megabaiti. Olemasolevat vastuse ekstraheerijat on uues süsteemis kavas kasutada koos tavalise veebipõhise infootsisüsteemiga (võimalike kandidaatide hulgast on praegu välja valitud Prise): see leiab tekstikogumikust dokumentide alamhulga, kus sisaldub küsijat huvitav informatsioon, millest vastuste ekstraheerimise blokk (ExtrAns) omakorda filtreerib relevantsed laused. Joonis 3 illustreerib süsteemi WebExtrAns töötamise põhimõtet: vastuste ekstraheerija on integreeritud infootsisüsteemi.



Joonis 3. Vastuse ekstraheerija integreerituna infosüsteemi

Uuele ainevaldkonnale üleminek tähendab ühtlasi uute keeleressurside – tesauruse ja terminite sõnastiku loomist, mis on Tartu rühma põhiülesandeks. Järgnevas anname ülevaate sellest tööst.

3. Link Grammar

Link Grammar ehk LG (http://bobo.link.cs.cmu.edu/link/index.html) on Carnegie Melloni ülikoolis väljatöötatud formalism, milles lause süntaktiline struktuur esitatakse niinimetatud seoste hulgana (*linkage*).

Selleks, et mõista LG põhiideed, vaatleme alljärgnevat näidet.

```
                    +----Os----+
 +-Dmc-+---Sp--+       +-Ds-+
 |     |       |       |    |
the dogs.n chased.v the cat.n
```

Toodud näites esinevad D-, S- ning O-tüüpi seosed. Igaüks neist ühendab täpselt kahte sõna (vasakpoolset ja parempoolset sõna):

1) *the* ja *dog* on ühendatud D-seosega, mis ühendab artikleid nimisõnadega;

2) *dog* ja *chased* on ühendatud S-seosega, mis ühendab nimisõnu finiitsete verbidega;
3) *chased* ja *cat* on ühendatud O-seosega, mis ühendab transitiivseid verbe oma laienditega.

Seosetüüpi täpsustavad väiketähed (toodud näites `s`, `p`, `m` ja `c`). Näiteks `s` ja `p` näitavad nimisõna–verbi ühildumist (vastavalt ainsus ja mitmus), `c` märgib loendatavust.

LG olulisim komponent on sõnastik, mis lisaks sõnadele toob ära ka sõnade ühendumist kirjeldavad kitsendused (nt nimisõna *dog* tohib olla D-seose paremal pool ja S-seose vasakul pool, aga mitte O-seose vasakul pool). Reeglistikus on esitatud sõna koos seostega, mida ta võib luua kas vasakule (–) või paremale (+).

Sõna	Seosed
a, the	D+
dog, cat	{@A-} & D- & {B+} & (O- or S+)
chased, bit	S- & (O+ or B-)
ran	S-
big, green, black	A+
Mary	O- or S+

Nagu tabelist nähtub, on kasutusel ka muud märgid peale +, - ja seosetähise. Need on selleks, et määrata kindlaks seose esinemise tingimusi. Mõned sõnad võivad teatavatel tingimustel anda ühesuguseid seoseid, kuid mitte teistsuguseid, samuti on tähistatud see, kas mingi seose esinemine on kohustuslik või võimalik.

LG-s on suuremad sõnagrupid koondatud nn. *words*-failidesse, seostekomplekt antakse kogu faili kohta. Nii näiteks moodustavad omaette failid transitiivsed verbid, intransitiivsed verbid, loendatavad nimisõnad, loendumatud nimisõnad.

Lause on LG tähenduses korrektne juhul, kui leidub seoste hulk, millega on täidetud järgnevad tingimused:
1) kõik sõnad lauses rahuldavad omavahel ühendumist kirjeldavaid kitsendusi;
2) seosed ei ristu, kui need on joonistatud sõnade peale nagu eespool toodud näites (planaarsus);
3) kõik sõnad on mingil viisil seotud, st puuduvad nn saared (sidusus).

Seega püüab *Link Grammarit* kasutav süntaksianalüsaator igale vaadeldavale lausele kirjeldatud tingimusi täitvat seostehulka vastavusse seada. Juhul kui see ülesanne õnnestub, on lause *Link Grammari* terminites süntaktiliselt korrektne. Lause võib olla ka LG jaoks mitmene – juhul, kui leidub rohkem kui üks tingimusi täitev seoste hulk. Sellise olukorra tekitab näiteks asjaolu, kus ühte sõna on võimalik tõlgendada nii verbi kui noomenina ja mõlemal puhul on seostetingimused täidetud.

4. Lihtsustatud inglise keel

Erivajadusteks kasutatavaid keeli (*language for special purposes*, LSP) iseloomustab lihtsus ja eriline reeglipärasus, nad on loodud mingis spetsiifilises ainevaldkonnas kasutamiseks. Vältida tuleb mitmetimõistetavust ja metafoore, kasutatakse eelnevalt defineeritud või üheseltmõistetavaid termineid, eelistatakse teatud verbivorme ja lausekonstruktsioone. LSP rangem juhtum on kontrollitud keeled, milliseid on loodud juba 1930. aastatest alates. Kontrollitud keelte loomise motiiviideks on tehnika intensiivne areng ja asjaolu, et tehnilisest tekstist peab hõlpsalt ja üheselt aru saama ka inimene, kelle emakeeleks pole keel, milles tekst on kirjutatud, kuid kes on spetsialist alal, millest tekst räägib.

Kontrollitud keele lihtne ja piiratud vorm loob eeldused selle kasutamiseks ka keeletehnoloogia rakendustes. Seni pole kontrollitud keeli siiski kuigi palju, selliseid süsteeme on kirjeldanud Nyberg ja Mitamura (1996) ning Knops ja Depoortere (1998).

AECMA (*Aircraft European Contractors Manufacturers Association*) lihtsustatud inglise keel (*Simplified English* ehk SE, http://www.aecma.org/sebrc.htm) on üks kontrollitud keeltest ja koosneb standardiseeritud üldsõnastikust ja komplektist reeglitest. SE sõnastikus on ligikaudu 900 sõna ja umbes 55 grammatikareeglit (need arvud võivad versiooniti veidi erineda) (Farrington 1996).

Kuigi SEl on oma sõnastik, ei tähenda see, nagu ei tohiks kasutada muid sõnu kui vaid need, mis selles üles loetud. Piirdudes vaevalt tuhande sõnaga, poleks võimalik kirjutada midagi nii keerulist nagu lennukimanuaal. SE juhend annab võimaluse kasutada manuaali kirjutamisel kolme liiki sõnu:

1) sõnad, mis on SE poolt tunnustatud ja sõnastikus loetletud – lisaks sõnakasutuse piirangutele tohivad ka lubatud sõnad esineda vaid ühes tähenduses, samuti tohib kasutada vaid neid vorme, mis on sõnastikus märgitud.
2) sõnad, mis kvalifitseeruvad tehniliste nimetustena (nimisõnad ja adjektiivid);
3) sõnad, mis kirjeldavad tootmisprotsessi (verbid).

Kõige avaramat tõlgendust võimaldab tehniliste nimetuste kategooria. SE juhendis on selle alaliike üles loetud kakskümmend. Siia kuuluvad näiteks nii lennuki osade (õlifilter, mootor, indikaator, propeller) ja kohtade (tiib, kabiin, paneel) nimetused kui ka matemaatilised, füüsikalised ja inseneriteaduslikud terminid (raadius, koefitsient, energia, faas, kõvadus), samuti navigatsiooni, meditsiini ja paljut muud vajalikku puudutavad.

Paraku on tegelikkus tunduvalt mitmekesisem kui teooria seda ette näeb. Käsitletava manuaali kirjutajad pole sageli üldse SE reeglitest kinni pidanud, meile üleantud tekstis leidub koguni õigekirjavigu. Viimatimainitud asjaolu sundis meid esialgseid plaane muutma ning tekitas vajaduse luua ka õigekirjakontrolli moodul.

5. Tehnilise käsiraamatu tekst

Tekstikogumiks, millest teadmusbaas koostatakse ja millega meil esialgu töötada tuleb, on AIRBUS INDUSTRIE lennuki hoolduse käsiraamat ehk manuaal (*Aircraft Mainetenance Manual*, AMM). See käsiraamat on kirjutatud SGML (*Standard Generalized Mark-up Language*) märgendust kasutades ning põhineb Lennutranspordi Assotsiatsiooni (*Air Transport Association* ehk ATA) poolt väljaantud dokumendikirjeldusel (*Document Type Definition* ehk DTD). SGML-märgendus lihtsustab oluliselt teksti automaatset analüüsi, kuna see võimaldab üheselt määratleda teksti peatükke, alapunkte, hoiatusteateid, pealkirju, tabeleid, loendeid jne. Ilma märgenduseta oleks automaatne analüüs keerukam ja vigaderohkem, kuna sel juhul saaks oluliste tekstikomponentide algust ja lõppu ainult aimata. Eeltoodud põhjustel kasutataksegi tehniliste ja mahukate dokumentide loomisel SGML (või ka XML) märgendust. Lõppkasutajatele tarnitakse elektroonilist käsiraamatut kahes versioonis: SGML-märgendatud tekstina või spetsiaalsele MS Windowsi keskkonnas töötavale lehitsejale (brauserile) loetavasse formaati teisendatuna.

Meie kasutame oma tööks esimest varianti, mis on teisendatud XML kujule. XML kujutab endast SGML-i lihtsustatud ja reeglipärasemat dialekti.

Kirjeldatava käsiraamatu puhul on kahjuks tihti eksitud SGML-ideoloogia vastu. Ligikaudu neljandik tabelitest ning loenditest on manuaalis esitatud ainult inimesele loetaval kujul, st tabelites puudub märgendus ridade/veergude eristamiseks ning loendites loendi elementide eristamiseks. See muudab automaatse töötluse väga ebamugavaks ning seetõttu oleme loobunud vaid inimloetavate SGML-tsoonide analüüsimisest.

5.1. Õigekirjavead

Käesolevas projektis on üks Tartu osapoole ülesandeid ka terminite leidmine. Paradoksaalsel kombel on terminite leidmine üsna tihedalt seotud õigekirjavigade leidmisega. Kuna termini üheks tunnuseks on tema teema-spetsiifilisus, siis leidub teda suure tõenäosusega just erialases tekstis ja mitte üldtekstis, kuid kõikvõimalikud laiatarbe-õigekirjakontrollijad on häälestatud üldiste tekstide tarbeks.

Reeglina kontrollib õigekirjaprogramm, kas tekstis leiduv sõna on olemas ka tema sõnastikus. Kui seda pole, siis on kaks võimalust:
1) sõna on tõesti valesti kirjutatud;
2) sõna on õigesti kirjutatud, aga seda ei leidu õigekirjaprogrammi sõnastikus.

Kuna tegemist on väga spetsiifilise tekstiga, siis on ülimalt tõenäoline, et õigekirjaprogramm leiab lisaks kirjavigadele ka suure hulga erialatermineid. Vigu terminitest eraldada pole aga esialgu võimalik muul viisil kui ainult käsitsi. Siiski, iga kord, kui mõnest tekstiosast on termin(id) leitud, siis saab neid lisada õigekirjaprogrammi sõnastikku.

Põhimõtteliselt on õigekirjavigadega toimetulemiseks kaks võimalust:
1) moodustada uus, ilma vigadeta versioon tekstist;
2) püüda vigadega kuidagi "käigu pealt" toime tulla.

Mõlemal võimalusel on omad head ja halvad küljed: kui meil on olemas vigadeta versioon, pole vaja enam vigade pärast muretseda ja nii teksti kasutajal kui ka teadmusbaasil on olemas õige variant, kuid samas me kaotame originaalteksti. Teiselt poolt, kui me püüame "käigu pealt" vigu parandada, on meil alati kasutada originaal-

versioon, kuid süsteem muutub selle võrra keerulisemaks. Nagu enamasti, osutub praktiliselt otstarbekaks vahepealne variant: sisendis on originaaltekst, õigekirjakontrolli moodul parandab vead (millest on eelnevalt moodustatud loend) ja kasutaja võib saada vastuse oma päringule kas juba parandatud variandist või soovi korral ka originaalvariandist.

Kuna manuaali maht on suur (üle 120 MB, rohkem kui 6,6 miljonit sõna), polegi vigade suhtarv teab kui oluline – ligikaudu 0,1%, aga praktikas tähendas see siiski rohkem kui 6000 sõna ülevaatamist ja sorteerimist. Pole võimatu, et manuaalis leidub ka selliseid vigu, kus sõna ise on õige (esineb õigekirjaprogrammi sõnastikus), kuid teda on kasutatud vales kontekstis. Paraku puudub meil selliste vigade leidmiseks vajalik aeg, tööjõud ja pädevuski. See nõuaks juba teksti täiendavat toimetamist erialaspetsialisti poolt.

6. Keeleressursid

Tartu töögrupi põhiliseks ülesandeks on välja töötada lingvistilised ressursid WebExtrAnsi jaoks. Realiseeritud kujul peaks see töö hõlmama leksikone, tesaurust ja terminoloogiat, mida kasutab peamiselt *Link Grammar*. Kõiki kolme liiki lingvistilisi ressursse saab luua peamiselt kahel viisil:

1) kasutades väliseid leksikaalseid ressursse ja viies neis leiduva eesmärgile vastava lingvistilise informatsiooni ühtsesse formaati;
2) kasutades teksti ennast (AMM) leksikaalse informatsiooni allikana.

Õnneks on meil lisaks manuaalile veel kasutada nii osade kataloog (IPC – *Illustrated Parts Catalog*) kui hulk ATA väljaandeid, kus on juttu nii täiendavatest reeglitest, mida peaks olema silmas peetud manuaali kirjutamisel, kui seletatud suur osa termineid ja lühendeid. Valdkonnaspetsiifiliste allikatena oleme siiani kasutanud ATA väljaannetes leidunud leksikaalset informatsiooni, kuid see ei välista ka muude valdkonnaspetsiifiliste ressursside kasutamise võimalust, kui seda vaja peaks minema.

6.1. Leksikonid

Leksikonid käesolevas töös on lihtsad sõnaloendid filtritele nagu speller ja *Link Grammar*. Sõnad on varustatud napi lisainfoga, mis

viitab nende kategoriaalsele kuuluvusele sõnaliiki või mõnda muusse filtris defineeritud klassi (näiteks lühend, viga vms). Kõige täpsemini vastab **leksikoni** mõistele LG leksikon. Käesolevaks hetkeks on seda juba täiendatud LG leksikonist puudunud, ent SE sõnastikus lubatud 145 sõnavormiga. Esialgu sai LG leksikonist eemaldatud koguni 308 sõnavormi, mille kasutamist SE ei lubanud, ent hiljem, SE ebaefektiivsuse ilmnedes, loobusime sõnavormide eemaldamisest ja piirdusime vaid puuduvate lisamisega.

Leksikonide hulka kuuluvaks võib pidada ka õigekirjaprogrammi (Spell) sõnaloendit. Et saada üle selle piiratusest, täiendasime seda alguses WordNeti sõnadega, hiljem lisasime ka neid termineid, mis manuaali tekstist teatavate märgendite vahelt õnnestus leida.

6.2. Tesaurus

Tesauruse eesmärgiks on ühest küljest toetada päringu analüüsimist, teisalt aga teadmusbaasi loomist/tekitamist, et samatähenduslikud tekstiüksused viitaksid kõik samale asjale tegelikkuses. Näiteks võib mingi lennukiosa või tööoperatsioon olla viidatud terminilaadse nimetusega, lühendiga või hoopis talle omistatud viitenumbriga (*reference number*). Tesauruses moodustavad eelmainitud üksused sünohulga (*synonym set, synset*) ja asuvad ühes kirjes, mille põhikuju on järgmine:

kirje number
    tekstiüksus
        [definitsioon, näited kasutuse kohta][1]
        [informatsioon üksuse iseloomu (nt termin) ja allika (nt CDSS) kohta]
tekstiüksus
        [definitsioon, näited kasutuse kohta]
        [informatsioon üksuse iseloomu (näit. lühend) ja allika (nt SE) kohta]
tekstiüksus
        [definitsioon, näited kasutuse kohta]
        [informatsioon üksuse iseloomu (nt reference number) ja allika (nt AMM) kohta]
[semantilised suhted (kehtivad kõigi ühes kirjes olevate tekstiüksuste kohta)]
        [semantilise suhte nimi ja viidatava kirje 1. tekstiüksus]

Esialgsete plaanide kohaselt peaks WebExtrAns kasutama kahte tesaurust:

[1] [ ] tähistab fakultatiivseid üksusi.

- **üldkeele tesauruse** moodustab WordNet 1.6 (või järgnevad versioonid) (http://www.cogsci.princeton.edu/~wn/w3wn.html), mis on registreeritav vabavara;
- **valdkonnatesaurus** koostatatakse EuroWordNeti (EWN) formaadis (http://www.hum.uva.nl/~ewn/).

EWN formaadi kasuks WordNeti formaadi ees räägivad asjaolud:
1) andmebaasi on võimalik sisse kirjutada kõikide andmete päritolu;
2) andmete sisestamiseks ja vaatlemiseks on olemas kasutajaliides Polaris;
3) Polarise import/exportformaat on ka tekstina lihtsalt loetav.

Valdkonnaspetsiifiliste allikatena oleme kasutanud erinevates ATA väljaannetes leidunud sõnaseletusi, definitsioone, lühendeid seletavaid tekstiosi. Et viia leitud valdkonnaspetsiifilist ja terminoloogilist informatsiooni ühtsele kujule, tuli töötada välja iga allika andmete struktuurile vastav variant Polarise import/eksportformaadis kirjest. Saadud tesauruse tekstiüksuste loendit tuleb järgnevalt testida analüüsitava teksti ehk manuaali peal. Kui palju valdkonnaspetsiifilistes allikates leidunud terminoloogiast on kasutatud manuaali tekstis ja kui palju terminite ja lühendite seletusi tuleb muudest allikatest (ka manuaali tekstist endast) juurde otsida? Sellele küsimusele veel vastata ei oska, kuid kindlasti saab manuaali tekstis terminitena märkida tekstiüksused, mis on välistes allikates täpsemalt lahti seletatud. Siiski on ka väliste leksikaalsete ressursside kasutamisel oht, et informatsioon muutub mitmetitõlgendatavamaks kui vaja. Näiteks on (erinevates) välistes allikates antud ühele ja samale lühendile erinevad seletused, aga analüüsitavas tekstis on kasutatud ainult ühte neist.

6.3. Terminid

Eraldi **terminoloogiate** järele puudub vajadus, sest valdkonnaspetsiifiline tesaurus täidab selle koha niigi. Ometi, kasutades eespool kirjeldatud 2. meetodit (tekst ise kui leksikaalse informatsiooni allikas), saab leida üles olulise osa terminitest ning suunata need omakorda tesauruse kirjeteks, millele hakatakse täiendavat informatsiooni lisama.

Paljud SGML-märgendid muudavad terminite automaatse eraldamise väga lihtsaks, sisaldades juba valmis termineid. Näiteks on märgendatud tööriistade nimed märgendipaariga

<TOOLNAME></TOOLNAME>. (Kogu tööriista nimetus koos numbriga on märgendite <TED> ja </TED> vahel.) Näiteks demagnetiseerijad

```
<TED>
<TOOLNBR>DM05275A</TOOLNBR>
<TOOLNAME>DEMAGNETIZER – 110V/60HZ (DM05275A)</TOOLNAME>
</TED>
```

ja

```
<TED>
<TOOLNBR>DM05275B</TOOLNBR>
<TOOLNAME>DEMAGNETIZER – 220V/50HZ (DM05275B)</TOOLNAME>
</TED>
```

Nõnda saab ekstraheerida hulgaliselt termineid, kasutamata järgnevas alapunktis kirjeldatud nimisõnafraaside tuvastajat.

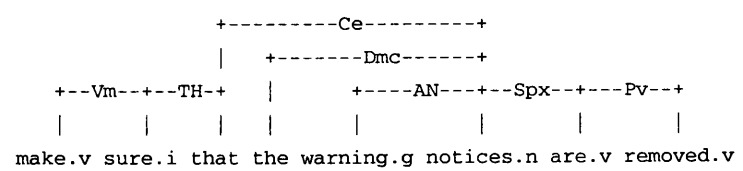
Mõistlik oleks rakendada tesaurust ja/või terminoloogiat *enne Link Grammari* süntaktilist analüüsi, et võimalikud terminid varakult üles leida ja päästa LG neile (või nende osadele) lahendust otsimast.

6.3.1. *Link Grammar* ja nimisõnafraaside tuvastamine

Kuna WebExtrAns kasutab LG formalismi oma sisendi süntaktiliseks analüüsiks, siis otsustasime kasutada seda ka terminite automaatseks ekstraheerimiseks.

Link Grammar on selles mõttes küllalt mugav formalism. Eeldusel, et *Link Grammar*il õnnestub lauset edukalt analüüsida, muutub ka nimisõnafraaside tuvastamine elementaarseks. Piisab vaid teatud seoste jälitamisest.

Vaatleme järgnevat, juba käsiraamatust võetud näidet:



Me teame, et S-tüüpi seose vasakpoolne sõna on nimisõna, mida too seos verbidega ühendab. Selleks, et kätte saada kogu nimisõnafraasi *the warning notices*, piisab, kui kogume kokku kõik alates sõnast *notices* saavutatavad sõnad (antud juhul laiend *warning* ning artikkel *the*).

Siiski leiduvad mõned seosed, mida mööda edasi liikuda ei tohi. Antud juhul on selliseks seoseks C-seos, mida mööda liikumine määraks nimisõnafraasiks *make sure that the warning notices*, mis on aga vale.

Üldine nimisõnafraasi definitsioon LG terminites võiks olla järgmine: nimisõnafraas sisaldab iga sõna, mis on saavutatav alustades seose x paremalt/vasakult poolt, vältides oma teel teatud seoseid.

Seega piisab, kui määratleme seosed x, mille parem (vasak) pool sisaldab nimisõnafraasi põhja, sellest lähtuvaid seoseid tulebki jälitada. Lisaks tuleb määratleda kitsendused seoste näol, mida tuleks vältida.

Selline definitsioon määrab maksimaalse pikkusega nimisõnafraasid, mis on moodustatavad igast lauses esinevast nimisõnafraasi põhjast. Maksimaalsed nimisõnafraasid ei ole aga lause ainukesed nimisõnafraasid, lisaks saab fraasist leida alamhulki, mis samuti nimisõnafraasideks kvalifitseeruvad.

Sellist teed on läinud näiteks firma Lingsoft (http://www.lingsoft.fi) nimisõnafraaside tuvastajaga NPtool, mis püüab nimisõnafraasi esitada kui kombinatsiooni eelmodifitseerijatest (*premodifier*), fraasipõhjast ning järelmodifitseerijatest (*postmodifier*). Eelmodifitseerijad on üldiselt adjektiivid ning artiklid, järelmodifitseerijad aga prepositsioonifraasid.

Valikuliselt fraasi komponente ära jättes saab näiteks fraasist *exact form of the correct theory of quantum gravity*, kus fraasipõhjaks on sõna *form*, moodustada järgnevaid alamhulki (vt http://www.lingsoft.fi/doc.nptool/term-extraction.html):

exact form of the correct theory of quantum gravity
exact form of the correct theory
exact form
form
form of the correct theory of quantum gravity
form of the correct theory

Lisaks leiab NPtool nimisõnafraasid, mis esinevad järelmodifitseerijates:

correct theory of quantum gravity
correct theory
theory of quantum gravity
theory
quantum gravity
gravity

Meie nimisõnafraaside ekstraheerija seni veel alamhulkade leidmisega ei tegele. Eelnevast lausest ekstraheerib ta fraasid:

> exact form of the correct theory of quantum gravity
> the correct theory of quantum gravity
> quantum gravity

st leitakse kõik maksimaalsed nimisõnafraasid, mis on saadavad põhjadest *form*, *theory* ja *gravity*.

Kuna meie töö eesmärk pole niivõrd nimisõnafraaside kui just terminite ekstraheerimine, siis on toodud lähenemist lisaks mõnevõrra täpsustatud.

Enamus nimisõnafraase ei kvalifitseeru terminitena. Küll võivad nad aga sisaldada termineid. Tihti õnnestub nimisõnafraasist teatud tüüpi sõnu välja filtreerida nõnda, et alles jääb vaid otsitav termin. Näiteks võib kustutada fraasi algusest artikleid ning asesõnu nagu *the, this, those* ja üldotstarbelisi adjektiive nagu *simple, old, particular, some* jne.

Nõnda leitud fraase võib veel töödelda ja näiteks analüüsida nende esinemissagedust tekstis. Sage esinemine on heaks kriteeriumiks fraasi tõlgendamisel terminina.

7. Kokkuvõte

Kogu projekti koordineerib Zürichi ülikool. Samuti on nende ülesandeks olemasoleva süsteemi ExtrAns täiustamine, sh dokumentide eeltöötluse mooduli ja mõningate vahemoodulite loomine, kogu süsteemi testimine ja võrdlemine standardsete infootsisüsteemidega. Tartu rühma ülesandeks on lisaks lingvistiliste ressursside loomisele ka semantilise ühestamise mooduli realiseerimine. Tööde koordineerimine toimub elektrooniliselt, selleks otstarbeks on loodud postiloend ja projekti veebileht (http://www.ifi.unizh.ch/CL/webextrans/), kus esitatakse mõlema rühma igakuised tööaruanded. Eesti arvutuslingvistidele on selles projektis osalemine hea väljakutse: kui senises rahvusvahelises koostöös oleme tegelnud eesti keele töötlemisega, siis siin on tegu kogemuse eksportimisega võõrasse keelde ja ainevaldkonda. Samas sunnivad uued ülesanded järjest juurde õppima ja pakuvad küllaga avastamisrõõmu.

Kirjandus

Farrington, G. 1996. AECMA Simplified English: an Overview of the International Aircraft Maintenance Language. – Proceedings of the First International Workshop on Controlled Language Applications. Leuven, Belgium: Katholieke Universiteit Leuven. 1–21.

Herzog, O., Rollinger, C.-R. (toim) 1991. Text Understanding in LILOG. Berlin, Heidelberg, New York: Springer Verlag.

Hess, M. 1997 Mixed-level knowledge representations and variable-depth inference in natural language processing. – International Journal on Artificial Intelligence Tools 6 (4), 481–509.

Hess, M. 1998. Antwortextraktion über beschränkten Bereichen. – Proceedings of KONVENS-98. Bonn. 337–346.

Knops, U., Depoortere, B. 1998. Controlled language and machine translation. – Proceedings of the Second International Workshop on Controlled Language Applications. Pittsburgh, Pennsylvania: Carnegie Mellon University. 42–50.

Mollá Aliod, D., Berri, J., Hess, M. 1998. A real world implementation of answer extraction. – Proceedings of 9th International Conference and Workshop on Database and Expert Systems. Workshop "Natural Language and Information Systems" Vienna. 143–148.

Nyberg 3rd, E. H., Mitamura, T. 1996. Controlled language and knowledge-based machine translation: principles and practice. – Proceedings of the First International Workshop on Controlled Language Applications, CLAW-96. Leuven, Belgium: Katholieke Universiteit Leuven. 74–83.

Sleator, D. D., Temperley, D. 1991. Parsing English with a Link Grammar. Technical report CMU-CS-91-196. Carnegie Mellon University, School of Computer Science.

Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse

Tiit Hennoste, Kadri Muischnek
Tartu Ülikool

Käesolev artikkel jaguneb kaheks osaks. Esimeses pooles anname ülevaate eesti kirjakeele korpusesüsteemi tekstivaliku põhimõtetest, teises osas vaatleme kahe valdkonna tekstide, ilukirjanduse ja ajakirjandutekstide sagedussõnastike algusosas toimunud muutusi 20. sajandil. Võrreldavaks on valitud perioodid 1930., 1960. ja 1990. aastad.

Eesti kirjakeele korpust on tehtud alates 1991. aastast. Tööd alustati eesti keele laboris ning jätkati sama seltskonnaga üldkeeleteaduse õppetooli juures olevas arvutuslingvistika töörühmas. Alguses juhtis korpuse tegemist Tiit Hennoste, alates 1995. aastast Heiki-Jaan Kaalep. Olulisemad rollid on jaotunud järgmiselt.

Korpuse tekstivaliku põhimõtted on välja töötanud ja tekstid valinud Tiit Hennoste. Korpuse tehnilist tegemist ning korpuse viimist arvutiformaati on juhtinud Heiki-Jaan Kaalep. Korpuse põhjal tehtud sagedussõnastikud on koostanud Leho Paldre ja Kadri Muischnek. Korpused on viinud internetti ning koostanud nende kasutamiseks vajalikud programmid Leho Paldre.

Korpuse tegemist on aegade jooksul rahastanud mitmed asutused ja fondid, millest keksed on ETF (grant nr 52, Eesti keele tekstide arvutikorpuse loomine (1991–1995)) ja Avatud Eesti Fond (korpuse paigutamine internetti ja kasutajaliidese tegemine, projekt Stylus 1996–97). Käesoleva artikli tekstide valikut kirjeldava osa on kirjutanud Tiit Hennoste, tekstide statistilised analüüsid on teinud Kadri Muischnek.

1. Eesti kirjakeele korpuse tekstivaliku põhimõtted

Eesti kirjakeele korpuse tegemist alustati baaskorpusest, mis oli analoogiline inglise keele klassikalise tekstikorpuse LOBiga ning millesse valiti tekste aastatest 1984–87/88 (vt Hennoste 1996; Hennoste jt 1998). Seejärel on koostatud ülejäänud allkorpused,

millest on seni valminud ainult ilukirjanduse ja ajakirjanduse tekstide osa. Korpus on praegu pooleli ja Internetis üleval järgmises seisus:

- baaskorpus (1984–87/8): ilukirjandus, ajakirjandus, populaarteadus, esseed ja biograafiad, hobi- ja harrastustekstid, propaganda, entsüklopeedilised tekstid, dokumendid, vaimulikud tekstid;
- ilukirjandustekstid aastatest 1890–1899, 1900–1910, 1911–1920 (praktiliselt 1917), 1935–1939, 1945–1954, 1966–1970, 1971–1975, 1988–1998;
- ajakirjandustekstid aastatest 1890–1899, 1900–1910, 1911–1920, 1935–1939, 1948–1952, 1966–1970, 1971–1975, 1988–1991, 1992–1995, 1996–1998.

Lisaks on Internetis olemas Projekti ELAN korpus (*European Language Activity Network* http://solaris3.ids-mannheim.de/elan/). Selle projekti käigus koguti veel umbes 1 miljon tekstisõna ajalehetekste, mis ajaliselt jaotuvad järgmiselt:

- ajakirjandustekstid 1985, 1986, 1989;
- ajakirjandustekstid 1990–1996;
- ajakirjandustekstid 1999.

ELANi eesmärgiks on luua rahvusvaheline standard kogudes standardiseeritud formaadis keeleressursse ja luues ühine päringukeel (ELAN-CQL) 27 Euroopa keele jaoks ning luua keeleressursside loojate ja kasutajate vastastikuse abi või ühistegevuse võrgustik.

ELAN-i viimast osa, 1999. aasta ajalehetekstide Internetist tõmbamist ja automaatset märgendamist võiks pidada pilootprojektiks käesoleval aastal alanud projektile "Eesti keele koondkorpuse loomine ja hooldamine" (rahastatakse riiklikust sihtprogrammist "Eesti keel ja rahvuskultuur"), mille sisuks on elektroonilisel kujul olevate tekstide kogumine ja ühtsesse formaati viimine võimalikult automatiseeritult.

Esialgu oli kavas lisaks baaskorpusele koostada samade valikuprintsiipidega korpused erinevatest eesti kultuuri ja keele perioodidest. Neid pidi siduma omavahel niitkorpus ilukirjandus- ja ajakirjandustekstidest valikkorpuste vahele jäänud aastatest (vt selle süsteemi kohta Hennoste 1996). Aja jooksul on kavad muutunud ning praegu on tegu korpusega, mis eesmärgina peab hõlmama eesti 20. sajandi avalike kirjalike tekstide põhirühmi kõigist perioodidest.

Baaskorpuse tekstivaliku põhimõtteid oleme kirjeldanud mitmes artiklis (Hennoste 1996, Hennoste jt 1998). Käesolevas artiklis lisame neile muude perioodide valminud korpuste tekstivaliku põhimõtted.

1.1. Tekstide valikut määravad situatiivsed variaablid

Tekstikorpus on tekstide kogum, mida iseloomustavad järgmised tunnused:

- ta on koostatud kindlatel eesmärkidel;
- eksplitsiitsete koostamisprintsiipide alusel;
- selleks, et iseloomustada keele mingit seisundit või varianti;
- ta on lõpliku suurusega kogum arvutiga töödeldavaid tekste;
- valitud nii, et ta oleks maksimaalselt representatiivne keelevariandi suhtes, mida ta peab esindama (Atkins jt 1992; Sinclair 1991; McEnery, Wilson 1997: 21–24).

Korpuste tegemisel tuleb teha valik ühiskonnas ringlevate tekstide hulgast. Lähtekohaks saab olla kas mingi hüpoteesitav allkeel või keelevälised kriteeriumid. Viimased võivad olla sellised, mis lähtuvad keelt mõjutavatest sotsiaalsetest ja situatiivsetest teguritest või sellised, mis seda ei arvesta. Klassikalised tekstikorpused olid kõik sellised, mis püüdsid arvestada keelt mõjutavaid keeleväliseid tegureid ning pakkuda tekstivalikut, kus oleks sobivas proportsioonis esindatud korpusekoostajate arvates olulised ja vajalikud tekstirühmad. Sellise lähenemise põhjused olid suuresti tehnilised: kuna korpusi sisestati käsitsi ning arvutite mälumahud olid väikesed, siis püüti saavutada võimalikult vähese tööga tulemust, mis oleks (all)keele suhtes maksimaalselt representatiivne.

Selline lähenemine on jäänud endiselt kehtima minevikutekstide korpuste tegemisel, mida tuleb käsitsi või skännides arvutisse sisestada. Tänapäeva tekstidest tehtavates korpustes sellist valimisviisi ei kasutata. Esiteks ei piira korpuste mahtu enam arvutite mälupiirangud. Teiseks on uurijad leidnud, et suure osa keeleprobleemide lahendamiseks on tarvis mitu suurusjärku suuremaid korpusi kui miljon sõna. Kolmandaks, uuema aja tekstid on tihti arvutites. Seetõttu on mindud üle automaatsele tekstikorjamisele kas internetist või kirjastuste arvutitest.

Selliste korpuste tekstivaliku põhimõtted on palju lõdvemad ja tekste valitakse suuresti selle alusel, mida on kergem kätte saada

ning arvuti abil automaatselt korpuseks töödelda. Praktiliselt on sellisel juhul tulemuseks üsna juhuslik tekstivalik, kui pole võimalik koguda sadu miljoneid tekstisõnu süstemaatiliselt nagu nt inglise keeles.

Sellised korpused on hästi kasutatavad keele universaalsete omaduste analüüsimiseks ning leksikograafiliseks tööks, kuid halvasti allkeelte erinevuste uurimiseks. Samuti on nad hästi kasutatavad kvalitatiivseks uurimistööks (nt maksimaalselt paljude sama sõna tähendusvariantide väljatoomiseks), kuid halvasti kvantitatiivseks uurimistööks. Põhjuseks on see, et eri aastatest on korpuses erineval hulgal ja erinevate kriteeriumide järgi valitud tekstikogumid. Muidugi võib uurija ise valida suurest korpusest oma alakorpused, kuid korpuse juhusliku valiku tõttu võib juhtuda, et võrdlemiseks sobivad alamkorpused on sama väikesed kui vanas stiilis valikkorpuse puhul (vt McEnery, Wilson 1997: 21–23)

EKKs on lähtutud keelt mõjutavatest keelevälistest mõjuritest nagu LOBis, kuid täpsustatud sealseid kriteeriume eesti kultuurisituatsioonile sobivaks.

EKK aluseks olevad tekstid on valitud järgmiste situatiivsete tunnuste alusel (vt ka Hennoste 1996, Hennoste jt 1998).

1. Sfäärid, milles keelt ühiskonnas kasutatakse, jaotatakse sotsiolingvistikas era/argi ja ametlikeks/avalikeks. Korpusse on võetud ainult ametlikus/avalikus sfääris ringelnud tekstid.
2. Igas ühiskonnas esineb nii emakeelseid tekste ning tõlketekste. Korpuses on üksnes emakeelsed tekstid.
3. Avalikus sfääris ringleb 3 tüüpi tekste: kirjalikud lugemiseks määratud tekstid, kirjalikud kuulamiseks määratud tekstid (raadiouudised jms) ning tekstid, millest pole kirjalikku fikseeringut ja mis on määratud kuulamiseks (raadiointervjuud jms). Korpusse on valitud ainult esimese rühma tekstid.
4. Sellised kirjalikud tekstid võivad esineda nii trükitud kui käsikirjalisel jms kujul. Me oleme valinud ainult trükitud tekstid.
5. Avalikus situatsioonis esinevad kirjalikud trükitud tekstid on valdavalt autori poolt ette valmistatud ja redigeeritud. Lisaks võib harva esineda ka spontaanseid tekste (nt internetiportaalide kommentaarid). Spontaansed tekstid on korpusest välja jäetud.
6. Tekstid on võimalik jagada luuleks ja proosaks. Korpusse on võetud ainult proosatekstid.

7. Avalikus situatsioonis tegutsevad erinevad tekstide autorid ja adressaadid. Me oleme sisse võtnud üksnes täiskasvanute poolt teistele täiskasvanutele kirjutatud tekstid.
8. Suuremal osal kultuuridest on olemas nii emamaa kui ka diasporaad (väliseestlaste kogukonnad Kaukaasias, Rootsis jm). Korpusesse on võetud ainult Eestimaal ringelnud tekstid
9. Kuna tekstide kirjutamisaega pole üldjuhul võimalik määrata, oleme valinud aluseks ilmumisaja. Teiseks, sama teksti võib avaldada mitu korda. Meie oleme valinud üksnes esmatrükid.

Saadud tekstikogumit on võimalik iseloomustada veel nelja parameetriga, mida pole valikul arvestatud, kuid mille väärtused on piisavalt suure tõenäosusega ennustatavad.

1. Tekstide ehitamine ja keelekasutus ühiskonnas on alati määratud teatud retooriliste ja keeleliste normingutega. Eesti ühiskonnas on 1930. aastatest nõutud, et avalikud tekstid järgiks õigekeelsusnorminguid, mis on kehtestatud selleks volitatud institutsioonide poolt. Normingute järgimist kontrolliti rangelt (ka. ilukirjandustekstid). See piirang lõdvenes taas 1990. aastatel.
2. Avalikus sfääris esinevad valdavalt kõrg- või keskharidusega inimesed.
3. Eesti kultuuris on alates 1930. aastatest eksisteerinud toimetajate ja korrektorite armee, kes avalikke tekste redigeeris. Seega peegeldab korpus oluliselt nende keelekasutust ja keeleideaale.
4. Kultuuris on oluline jaotus linna ja maa vahel nii kultuuriliselt/tekstiliselt (linnakultuur ja linnakultuuri tekstid *contra* maakultuur ja maakultuuri tekstid) kui ka keeleliselt (linnamurded/maamurded). Võib suure tõenäosusega väita, et korpuse tekstid esindavad linnakultuuri ja linnakeelt.

Eesti kultuuris iseloomustab eeltoodud kriteeriumidele vastavaid tekste see, et ülivaldavalt kasutatakse neis kirjakeelt ja need tekstid moodustavad kirjakeelsete tekstide tuumosa. Seega võime väga suure tõenäosusega väita, et tegu on eesti kirjakeele korpusega.

1.2. Tekstiklassid

Eelnevate kriteeriumide alusel valitud tekstid on korpuses jaotatud tekstiklassideks ja need omakorda alamklassideks. Igast teksti-

klassist on valitud korpusesse teatud hulk tekste. Tekstiklasside ja nende suuruste määramisel on aluseks LOBist üle võetud faktorid:
1) aine/valdkond;
2) stiil;
3) publitseerimise meedium ehk publitseerimisviis;
4) suhteline mõju või autoriteet;
5) trükiarv või levik.

Esimesed kaks faktorit annavad kokku tekstiklassid (teadus, ajakirjandus jms), mille aluseks on tegelikult raamatukogude kataloogide liigendus ja mis viitavad sellele, millises ühiskonnaelu valdkonnas tekstid ringlevad. See liigendus kattub vaid osaliselt tekstide grammatiliste ja leksikaalsete tunnuste alusel tehtud liigendusega (vt Douglas Biberi uurimusi inglise keele põhjal, Biber 1989). Publitseerimisviis eristab raamatud, kogumikud (+ perioodika) ning dokumendid. Suhteline mõju ja trükiarv annavad kokku retseptsiooniindeksi, mis on aluseks tekstide hulga üle otsustamisel. Kuna eesti kultuuris puuduvad uuringud, mis lubaks mõõta tekstide mõjukust, siis oleme võtnud aluseks ainult trükiarvu või leviku (vt pikemalt valimise kohta Hennoste jt 1993: 591–592).

Nii on nende faktorite põhjal saadud tekstide jaotus eesti kultuuri tarvis suhteliselt hästi ülevõetav. Erinevustest ja probleemidest tuleb juttu järgnevas osas.

Tekstide kogumaht kategooriates on laias laastus võetud üle LOB-ist, kuid sealseid proportsioone on muudetud sobivaks eesti tekstisituatsiooniga. LOBi eeskujul pidi korpus sisaldama 1 miljon tekstisõna, mis oli jagatud 500 tekstiks, igast üks väljavõte 2000+/-200 sõna. Tegelikus korpuse tegemise käigus tuli neist ideaalidest kõrvale kalduda. LOBis oli määratud võimalikult jäigalt üks tekstivalim 2000 sõna pikkuseks. Kui valitud tekstis polnud nii palju sõnu, siis lisati valimisse sama tekstirühma teisi tekste. Meie korpuses oli määratud tekstirühma kogusuurus (nt ilukirjandustekstide maht). Kui valitud tekstis polnud piisavalt sõnu, siis seda valimit ei täiendatud, vaid lisati tekstirühma tekste juurde.

1.3. EKK struktuur ja tekstivaliku alused

Kuna EKK koosneb praegu osadest, mis on erineva suuruse ja valikuga, siis iseloomustame eraldi baaskorpust ja muid korpusi.

EKK baaskorpuse struktuurist annab ülevaate Tabel 1. Sellest on näha erinevad tekstiklassid, aastad, millest on vastavate klasside tekstid valitud, tekstide ja neis olevate sõnade hulk ning see, mitu protsenti moodustab vastav tekstiklass kogukorpusest.

Tabel 1. EEK baaskorpuse struktuur

Kategooriad	Aastad	Tekstid	Sõnad	%
ABC **Ajakirjandus**	1985	519	176017	17,2
D **Religioosne kirjandus**	1984–6	4	8011	0,8
E **Hobid ja harrastused**			75410	7,4
Raamatud	1984–6	20	39572	
Perioodika	1984–6	45	35838	
F **Populaarkirjandus**			164218	16,1
Raamatud	1984–6	49	150024	
Perioodika	1985	37	14194	
G **Esseed ja biograafiad**			90661	8,9
Raamatud	1985, 5, 7	16	32017	
Perioodika	1984, 5	36	58644	
I **Mitmesugust**			12427	1,2
Dokumendid	1984–6	8	12427	
J **Teadus**			155448	15,2
Raamatud	1984–6	49	96235	
Perioodika	1985	37	59213	
KLMNPR **Ilukirjandus**			255416	25,0
Raamatud	1984–7	93	192667	
Perioodika	1984–7	35	62749	
S **Entsüklopeediad**	1984, 5, 8	11	22769	2,2
T **Propaganda**			60256	5,9
Raamatud	1984–6	14	28638	
Perioodika	1985	16	31618	
KOKKU	**1984–88**	**989**	**1020645**	**100**

Ülejäänud aastatest on esmalt valitud korpusesse ilukirjanduse ja ajalehtede tekstid seetõttu, et just nemad moodustavad eesti kultuuris tekstide põhimassi ning need on ka eesti kirjakeele näidete kogumise ja normingute kehtestamise põhikohad. Samal ajal esindavad nad kirjalike avalike allkeelte seas erinevaid registreid (vt inglise keele kohta Biber 1989).

Ilukirjandus (täpsemalt kujutluslik narratiiv) moodustab suhteliselt omaette seisva allkeele, mille tuumaks on realistlik ilukirjandus, mis on eesti kultuuris valitsenud läbi kogu 20. sajandi.

Ajakirjandus kuulub koos populaarteaduse, elulugude, praktiliste käsiraamatute jms kõige laiemalt levinud neutraalsesse allkeelde, mida D. Biber nimetab *general narrative exposition*. Seda allkeelt iseloomustab see, et temas pole grammatilisi ja leksikaalseid jooni, mis statistiliselt esile tõuseksid. Selliselt annab ta keelekasutuse, mida mitmetes teooriates nimetatakse normaalproosaks ning vaadeldakse kui kirjakeele tüüpilisimat esindajat.

Teiseks käituvad need kaks valdkonda ajas keeleliselt erinevalt. Ilukirjanduskeel sõltub konkreetsetest autoritest ning kirjandusvoolude ja suundade vaheldumisest, mis ei seostu kuigi palju ühiskonnas toimuvate muutustega. Vaid nõukogude süsteemi teke ja kadumine mõjutasid ilukirjanduse keelekasutust tugevamalt. Ühtlasi muutub ilukirjanduskeel suhteliselt aeglaselt. Ajakirjanduse keelekasutus on tugevalt seotud ühiskonnamuutustega ning muutub väga kiiresti, kuna peab vahetult kajastama ühiskonnas toimunud protsesse. Muud valdkonnad on 20. sajandit kui tervikut vaadates eesti kultuuris perifeersed.

Tabel 2. Ilukirjanduse ja ajakirjanduse korpuste struktuur 1890–1998

	KLMNPR **Ilukirjandus**	ABC **Ajakirjandus**
Aasta	**Sõnu**	**Sõnu**
1890–1899	155 000	193 000
1900–1910	64 500	171 500
1911–1920	247 000	182 500
1935–1939	252 000	117 000
1945–1954	66 000	–
1948–1952	–	242 400
1966–1970	132 000	201 000
1971–1975	257 100	168 500
1984–1987 (baas)	255 416	176 017
KOKKU	**1 428 916**	**1 451 917**
1988–1998 (TEI)	611 000	384 800

Asjaajamine (dokumendid) oli 19. sajandi lõpul ja 20. sajandi algul enne Eesti Vabariigi teket venekeelne. Samuti oli suur osa asjaajamisest venekeelne nõukogude perioodil.

Teadus oli 19. sajandi lõpul ja 20. sajandi algul valdavalt saksa- ja venekeelne. Samuti oli suur osa teadust venekeelne nõukogude perioodil. Viimastel aastatel on aga eriti reaalteadused olnud valdavalt ingliskeelsed. Samuti on olulisi erinevusi eri teadusalade

keeles. Enam on eestikeelset teadust olnud humanitaaraladel, reaalteadused on läbi olnud eelkõige võõrkeelsed. Samal ajal moodustavad teaduse ja asjajaamise keeled suhteliselt iseseisva ja neutralkeelest ning narratiivi keelest statistiliselt erineva allkeelte kogumi (vt Biber 1989), mille muutumine jääb praegu kahjuks korpuses kajastamata.

Kategooriad E, F ja G ehk populaarteadus, mitmesugused käsiraamatud (kokaraamatutest taimemäärajateni), biograafiad jms on statistiliselt laiemalt levinud kui teadus, kuid vajadus nende järele korpuses on väiksem, sest eeldatavasti kuuluvad nad ka eesti keeles eelkõige neutraalstiili nagu teistes uuritud keeltes.

Usuline kirjandus, propaganda ja entsüklopeediad on olnud 20. sajandi eesti kultuuris marginaalsed valdkonnad. Usulist kirjandust ilmus nõukogude ajal minimaalselt ja seegi ei saanud avalikku levikut. Entsüklopeediad jm teatmeteosed on olnud läbi sajandi harvad. Propagandakirjandus esines sellisel kujul ainult nõukogude perioodil. Tema tiraažid olid väga suured, kuid tegelik levik väga väike.

1.4. Tekstiklassisisesed valikuprobleemid

Järgnevalt vaatleme probleeme, mis tekkisid LOBi tekstijaotuse järgimisel ja meie pakutud lahendusi nendele probleemidele. Pikemalt peatume ajakirjanduse ja ilukirjanduse tekstide valikute juures. Muud tekstirühmad on seni olemas ainult baaskorpuses. Neis oli valikuprobleeme kolmes rühmas (vt ka Hennoste jt 1993: 593–594; Hennoste 1996).

J: Teadus. Teadustekstid on LOBis jaotatud vastavalt teadusaladele (loodusteadused, meditsiin, matemaatika, sotsiaalteadused, jne). Need valdkonnad on omakorda jaotatud kitsamateks teadusaladeks. EKKs on teadus võetud üheks tervikuks, kuna suur osa teadusaladest eesti kultuuris puudub või on neis tekste väga vähe.

Teiseks probleemiks on allikad ja nende levik. Nõukogude Eestis jagunesid teadustekste avaldavad allikad kahte rühma. Esimesse kuulusid teaduslikud allikad (monograafiad, artiklikogud, ülikoolide toimetiste sarjad jms). Teine osa teadustekste ilmus aga ajakirjades, mis ei kuulunud teaduslike ajakirjade hulka. Nt avaldati ajakirjas "Eesti Loodus" lisaks populaarteadusele igas numbris 1–2 puhtalt teaduslikku teksti. Sellepärast oleme valinud teadustekste mõlemast allikatüübist.

Kolmandaks, teadusalade klassifikatsioon on Eestis olnud erinev kui USAs või Lääne-Euroopas (nt lingvistika on Eestis traditsiooniliselt paigutatud humanitaarteaduste, mitte sotsiaalteaduste alla). Me oleme järginud korpuses Eesti traditsiooni.

S: Entsüklopeediad ja teatmikud. Entsüklopeediad ja teatmikud puuduvad LOBis eraldi kategooriana. Meil on nad paigutatud eraldi eelkõige kahel kaalutlusel. Esiteks, nad kuuluvad populaarteaduse ja teaduse vahele ja teiseks, nad moodustavad tekstiehituslikult omaette rühma, mida iseloomustab eelkõige kompressioon, mis tuleneb vajadusest mahutada paberipinnale maksimaalne kogus infot.

T: Propaganda. Omaette probleem on nõukogude-aegse tekstirühmaga, mille sisuks on marksism-leninism ja funktsiooniks riigi ametliku ideoloogia ja filosoofia propaganda. See rühm võeti raamatukogukataloogides ning ka muudes tekstide loendites eraldi ning tal on ka omad tekstiehituslikud tunnused, mis johtuvad eeskätt marksismiklassikute tekstilistest seisukohtadest. Selliseid tekste oli palju, kuid nende tegelik mõju väga väike, sest neid tavaliselt ei loetud.

Tekstirühmade valimisel on kasutatud perioodilisi väljaandeid "Raamatukroonika" ja "Artiklite ja retsensioonide kroonika" mis sisaldavad andmeid vastavalt ilmunud raamatute ning kogumikes, ajakirjades ja ajalehtedes ilmunud tekstide kohta. Religioonitekstide valimisel on toetutud ka Tartu ülikooli raamatukogu kataloogidele.

KLMNPR: Ilukirjandus. LOBis on ilukirjanduse osas eraldi välja toodud kommertsilukirjanduse erinevad alaliigid (L–R) ning jääk (K), kuhu kuulub nn tõsine ilukirjandus. Kuigi Eestis on ilmunud ajaviitekirjandusse liigitatavaid teoseid nii olulistelt autoritelt (E. Vilde, O. Luts, M. Traat jt) kui ka poolamatööridelt, puudub meil traditsioon liigendada kirjandust kaheks ning puudub ka järjepidevus ajaviitekirjanduse tüüpides (kriminullid, armastusromaanid jne). Seetõttu jätsime kogu kirjanduse üheks tekstiklassiks. Välja on jäetud luule, draama, lastekirjandus ja suurem osa huumorist. Esimesed kolm jäävad väljapoole meie pandud korpusepiire. Huumor kuulub eesti traditsioonis ajakirjandusliku publitsistika alla.

Kõikides kirjanduse allkorpustes on tekstivaliku põhimõtted ühesugused.

1. Eesti algupärase proosa kogutoodang on nii väike, et korpusse on valitud üks katke igast eesti keeles ilmunud proosaraamatust.

Romaani puhul on see antud romaani katkend, novellikogust on valitud üks katkend kogu kohta. Raamatute tiraažide erinevust pole seetõttu arvesse võetud. Arvesse on võetud ainult vastaval perioodil esmatrükis ilmunud teosed. Kordustrükid ja klassika uustrükid on välja jäetud.

2. Lisaks raamatutele on kirjandust ilmunud ka ajakirjanduses ja kogumikes. 19. sajandi lõpus ja 20. sajandi alguses ilmus ajalehtede lisasid, mis täitsid ajakirja rolli ja milles ilmus ka ilukirjandust. Samuti on läbi sajandi ilmunud ilukirjandust järjejutuna ajalehtedes. Valdavalt on see olnud tõlkekirjandus, kuid eriti 19. sajandi lõpus ilmus ka algupäraseid romaane (nt Eduard Vilde). 20. sajandil ilmus mitmeid üldhuviajakirju, milles kirjandus oli ainult üks, perifeerne osa. Kuna vähegi väärtuslik osa sellisest kirjandusest ilmus ka raamatutena, siis on ajalehed ja sellised ajakirjad korpusest välja jäetud. Välja on jäetud üksikud jutud, mis on ilmunud kalendrites ja mitmesugustes tähtpäevade puhul ilmutatud segakogumikes.

1923 hakkas ilmuma kirjanduajakiri "Looming", mis on väikese vaheajaga II maailmasõja ajal ilmunud tänini. See on pidev foorum, kus eelkõige ilmus novelle, aga ka romaanikatkeid või terveid romaane. "Looming" kui järjepidev ja autoriteetne kirjandustekstide allikas on korpusesse sisse võetud.

Lisaks "Loomingule" on olnud kaks mõjukat kirjandust avaldanud ajakirja. Esimene oli nõukogudeaegne noorteajakiri "Noorus", mis avaldas nooremate autorite töid. Palju ilmus neid 1960. aastatel, mil ajakiri oli ka eriti mõjukas kirjandusfoorum. 1986 loodi noorte kirjandusjakiri "Vikerkaar" See on avaldanud järjekindlalt nooremate kirjanike ilukirjandust ja olnud mõjukas tänaseni. Seetõttu on "Nooruse" 1960. aastate proosa ja "Vikerkaare" proosa ka korpusesse sisse võetud.

Ajakirjade osakaal koguvalikust on määratud võrreldes ajakirjade tiraaže raamatute kogutiraažiga: nad on saanud oma tiraažiga proportsionaalse osa mahust. Ajakirjadest on samuti valitud ainult uusi originaalteoseid. Kui valik sattus kokku juba raamatust valitud tekstiga, siis vastavat teksti kaks korda ei valitud.

3. Igast proosaraamatust on võetud korpusesse katke pikkusega 2000 tekstisõna, mis on valitud juhuvaliku põhimõttel raamatu algusest, lõpust või mingist konkreetsest leheküljest alates. Katke algab lühijutu, romaani peatüki või lõigu algusest ning lõpeb lõigu lõpuga. St valikud ei ole täpselt 2000 sõna pikad vaid arvestavad lõigupiire.

4. Ilukirjandustekstide ilmunud nimetuste leidmisel on kasutatud eri aegade kohta erinevaid allikaid.

19. sajandi materjalide aluseks on Eesti Kirjandusmuuseumi süstemaatiline kataloog, milles on toodud välja eesti romaanid, jutustused ja jutukogud pealkirjade järgi, bibliograafiad ajalehes Postimees (1901 nr 49 ja 51) ning kirjanike biograafiaid sisaldava koguteose "Eesti kirjarahva leksikon" (Kruus 1995) andmed. Praegune valik sisaldab ainult aastatel 1890–99 eraldi raamatutena ilmunud töid. Bibliograafilised materjalid ei garanteeri, et oleks kirjas kogu noil aastatel ilmunud uus algupärane proosa. Siiski moodustab see ülivaldava osa ilmunust ning garanteerib representatiivse valiku.

1900–1920. ilukirjanduse valiku aluseks on neil aastatel ilmunud ilukirjanduse kartoteek Eesti Kirjandusmuuseumis.

1935–1939. aastate ilukirjanduse valiku aluseks on "Eesti raamatute üldnimestik" mille 4. osa hõlmab aastaid 1934–36 (ilmunud aastal 1938) ja 5. osa aastaid 1937–39 (ilmunud 1940).

Nõukogudeaegsete ilukirjandustekstide ja 1990. aastate ilukirjandustekstide valiku aluseks on bibliograafiline ülevaade "Raamatukroonika" / "Eesti Rahvusbibliograafia" Selle tegemine katkes 1990. aastate alguses. Hiljem on kroonikat tehtud tagantjärele, kuid kõik pole veel valmis. Seetõttu puuduvad korpusest praegu 1993 ilmunud raamatud. Samuti puudvad 1998. aasta raamatud, sest selle aasta biblio polnud valiku tegemise ajal veel kättesaadav.

5. Proosavaliku aastad erinevad ajakirjandusvaliku aastatest kahel perioodil. 2. maailmasõja järgsetel aastatel ilmus palju ajalehti, kuid väga vähe raamatuid (1948–52 ilmus 16 algupärast proosateost). Seetõttu laiendati ilukirjanduse valikut aastatele 1945–1954. 1990. aastad on ajakirjanduses jagatud kolmeks alamperioodiks, sest nende perioodide ajakirjanduspilt erineb üksteisest väga tugevasti. Ilukirjanduses selliseid selgeid perioode ei ole, seetõttu on kogu ilukirjandus võetud üheks perioodiks.

ABC: Ajalehed. Ajalehtede valiku tegemisel ei ole võimalik kasutada samu valikupõhimõtteid kõigi perioodide kohta, sest lehtede süsteem on eri perioodidel olnud erinev. Lehtede valiku aluseks on olnud järgmised põhimõtted:

1. On valitud kaks lehte, mis on ilmunud läbi kogu 20. sajandi: põhiliselt Tartus ilmunud Postimees (nõukogude ajal Edasi) ja Viljandi kohalik leht Sakala (Punane Täht).

2. Levikult jagatakse lehed üleriigilisteks, suurlinnade/regioonide ja kohalikeks lehtedeks. Suurlinnalehtede lehtede hulka kuuluvad Tallinna ja Tartu lehed. Kohalikud lehed on sealjuures olnud eri aegadel erineva vormiga. 19. sajandil oli tegu eraldi lehtedega (nt Pärnu Postimees), mis ilmusid vastavas kohas ja levisid väljaandmiskoha lähiümbruses. 20. sajandi alguses levisid nende asemel enamasti suurte lehtede kohalikud väljaanded (Vaba Maa Pärnus), mis sisaldasid lisaks üleriigilisele osale ka kohalikke uudiseid. Selline süsteem kestis kuni 2. maailmasõjani. Nõukogude ajal loodi uus kohalike lehtede süsteem, mis on põhiosas säilinud tänapäevani. Selles süsteemis on igal rajoonil/maakonnal oma ajaleht.

1890. aastate ja 20. sajandi alguse lehed jagunevad kaheks: regionaalsed ehk Tallinna ja Tartu suured lehed (Valgus, Postimees, Olevik jms) ning kohalikud lehed (väikelinnade lehed, suurte linnade pisilehed ja üleriigiliste lehtede kohalikud väljaanded).

1930. aastatest alates jagunevad lehed kolme rühma: üleriigiliselt levivad lehed (1930. aastatel Kaja, Päevaleht jms, nõukogude ajal Rahva Hääl, Noorte Hääl jms), suurte linnade lehed (nt Postimees, Õhtuleht) ning kohalikud lehed ja üleriigiliste lehtede kohalikud väljaanded (nt Vaba Maa Pärnus, Pärnu Kommunist jms).

Lisaks on eri perioodidel ilmunud palju pisilehti (kolhooside lehed, alevite lehed, koolilehed jms). Need on jäetud korpusest välja, kuna nende hulk, nimetused, levik jms on enamasti teadmata. Igal juhul oli nende levik ja mõju väga väike.

3. Sisult jagunevad lehed kolme suurde rühma: kvaliteetüldlehed, kollased üldlehed (tabloidid) ja erilehed. Üldlehed kirjutavad paljudel eri teemadel (poliitika, majandus, sport jne). Erilehed on kas konkreetsele lugejarühmale määratud või kitsamat valdkonda kajastavad lehed (kultuurilehed, ärilehed jms). Kohalikud ja suurlinnade lehed olnud eelkõige üldlehed ning erilehed olnud eelkõige üleriigilised.

19. sajandi lõpu ja 20. sajandi alguse lehed olid kõik üldlehed. 1930. aastatel saab lehed jagada kolme rühma: kvaliteetüldlehed, kollased lehed (Esmaspäev, Tallinn Post) ja erilehed (Ühistegelised Uudised, Õpetajate Leht jms). Nõukogude aja lehed jagunesid kaheks: üldlehed ja erilehed (Sirp ja Vasar, Spordileht jms).

4. Ilmumissageduselt on lehed jagatud kolmeks: 5–7 korda nädalas ilmunud päevalehed, 2–3 korda nädalas ilmunud lehed ja nädalalehed.

19. sajandi lõpu lehed olid peaaegu kõik nädalalehed, ainus päevaleht oli Postimees. 20. sajandi lehed on jagunenud kõigi kolme rühma vahel. Sealjuures on üldlehed olnud valdavalt päevalehed ja harvemini ilmuvad lehed põhiliselt erilehed. Üleriigilised ja suurlinnade lehed on olnud valdavalt päevalehed, kohalikud lehed ilmunud kas nädalalehtedena või 2–3 korda nädalas.

5. Lehed võivad olla kindla poliitilise partei häälekandjad, kitsapiirilise kindla ideoloogia kandjad või rahateenimislehed, mille ideoloogiline platvorm on lai. 19. sajandi lehtede juures pole ideoloogiline jagunemine oluline. 20. sajandi alguse poliitiliste võitluste ajal jagunesid lehed ideoloogiliselt kahte selgesse rühma. Ühele poole jäi radikaalne rinne (Tallinnas Teataja 1901–05, Tartus Meie Aastasada 1911–16 jms) koos radikaalpahempoolse töölisajakirjandusega, mis tekkis 1905 aasta paiku. Teise poole moodustas konservatiivne suund (Postimees jms). Seda jagunemist on valikul arvestatud.

Eesti Vabariigi ajal oli valdav osa lehti seotud parteidega, kuid selget kaheksjagunemist ei olnud. Sama kehtib nõukogude aja kohta, mil lehtede parteilisus pole arusaadavatel põhjustel oluline. 1990. aastatel on kõik valitud lehed olnud rahateenimislehed.

6. Ajaleht sisaldab väga erinevat materjali: uudiseid, juhtkirju, arvustusi, infot, reklaami jne. See materjal jagatakse kahte suurde rühma: toimetuse materjal ja kuulutused+reklaam. Korpusesse on võetud ainult toimetuse materjal. Osa sellest on toimetuse enda kirjutatud, osa saadud teabeagentuuridelt ning tõlgitud teistest keeltest. Korpusesse on võetud originaalmaterjal ning Eesti teabeagentuuridelt saadud uudised, välja on jäetud tõlked.

LOBis on lisaks valiku aluseks erinevad lehežanrid, kuna neil on erinev funktsioon, ehitus ja keelekasutus. Läbi sajandi ulatuva valiku puhul pole sellist jaotust võimalik teha. Esiteks, žanrid kujunevad välja 20. sajandi jooksul ning ilmuvad lehtedesse eri aegadel. Teiseks, žanrikaanonid muutuvad läbi sajandi. Kolmandaks, ideloogia toob omapoolseid nõudeid žanriehitusse, eriti nõukogude ideoloogia. Seetõttu on alati viidud sisse kõik originaaltekstid, mis antud lehenumbris sisalduvad.

7. Lehtedest võetava materjali hulga otsustamisel on LOBis arvestatud retseptsiooniindeksit (levikut ja mõju). Kuna meil puudusid uuringud mõju kohta, siis oleme arvestanud eelkõige levikut ja

mõju üksnes lisakriteeriumina. Lehtedest võetavate valimite suuruste määramisel on lähtutud järgmistest põhimõtetest:

• alguses on otsustatud, mitu lehenumbrit peaks korpusesse kuuluma, et saada soovitud korpusemahtu;

• iga lehe valimi suuruse määramise aluseks on tiraaž. Kuna lehed on erineva ilmumissagedusega, siis on aluseks võetud arvestustiraaž. Selleks on arvutatud vastava lehe keskmine tiraaž antud aastatel. Sellega on korrutatud aastate arv, mille jooksul ilmus sellel perioodil ja päevade arv nädalas, mil leht ilmus (nt Päevaleht: 40 000 (aastate keskmine tiraaž) x 5 (ilmumisaastate hulk vastavas perioodis) x 7 (ilmumispäevi nädalas) = 1 400 000). Järgnevalt on arvutatud igale leherühmale kuuluva tiraažiprotsent lehtede koguhulgast, mille põhjal sai iga rühm vastava hulga lehenumbreid.

8. Eri perioodidel on kasutada olnud eri täpsusega allikad. Varasema ajakirjanduse nimestik on olemas R. Antiku biblios "Eesti ajakirjandus 1766–1930" nõukogude aja kohta on andmed "Raamatukroonikas"

Tiraažide kohta on 19. sajandi lõpu põhiallikaks varasema ajakirjanduse ajalugu "Eesti ajakirjanduse teed ja ristteed" (Peegel jt 1994). Tolle aja lehtede leviku kohta on andmeid väga vähe.

20. sajandi alguse kohta pole korralikku ülevaadet. Põhiallikaks on bibliograafia "Eesti ajakirjandus 1900–30" Kirjandusmuuseumis. Ajakirjanduse ajaloos oli see üleminekuperiood, mil suur osa eelmise perioodi lehti lõpetas ning alustasid paljud uued lehed. Neist aga jäi suure osa ilmumisaeg väga lühikeseks. Tiraažide kohta on tol perioodil väga vähe teada.

Eesti Vabariigi perioodi kohta pole samuti korralikku ülevaadet. Põhiallikaks on Helen Kulpa käsikirjaline biblio "Eesti ajakirjandus 1931–40" Kirjandusmuuseumis ja Epp Laugu artikkel "Eesti Vabariigi ajakirjandusest 1920.–1930. aastatel" (Lauk 1991). Korralikku ülevaadet selle aja lehtede tiraažidest olemas ei ole. Ka on tiraažid väga kõikuvad aastate ning isegi numbrite lõikes. Kohalike lehtede tiraažid ei ole teada. Võtsime lähteks 3000, mis on saadud E. Laugu tabelis olevate kohalike lehtede keskmiste põhjal.

Nõukogude Eesti lehtede kohta on andmed võetud "Raamatukroonikast"

9. Kolm probleemi oli Nõukogude Eesti lehtedega. "Kodumaa" oli määratud levitamiseks eelkõige väliseestlaste hulgas, kuid teda müüdi vabalt ka Eestis. Seetõttu võtsime ta korpusesse sisse.

Uudised ETAs (Eesti Telegraafi Agentuur) kirjutati osalt vene keeles ning tõlgiti peale tsensuuri tagasi eesti keelde. Praktiliselt võimatu on määrata, millal on tegu tõlkega, millal originaaliga. Seetõttu oleme kõik ETA uudised sisse võtnud.

Läbi kogu nõukogude aja on valitud sama lehekomplekt välja arvatud kaks lehte. Sõjajärgsest perioodist on valitud üheks erileheks Talurahvaleht ja hilisemast perioodist Kodumaa.

10. Teise probleemse perioodi moodustab taasiseseisvumine ja postsotsialistlik aeg (1988–98). See periood on väga kirju ning on seetõttu jagatud kolmeks alaperioodiks: perestroika (1988–91), segaduste aeg (1991–95) ja stabiilsuse algus (1995–98).

1. Läbi kõigi alaperioodide on valitud lehed, mis ilmusid edasi: Edasi/Postimees, Sirp ja Vasar/Kultuurileht/Reede/Sirp, Punane Täht/Sakala ning Õhtuleht. Kodumaa kaotas oma tähtsuse, lõpetas ilmumise 1991 ja teda pole selle perioodi valikutesse võetud.

2. On toodud sisse uusi lehti, mis püsisid ja olid mõjukad. Sealjuures on valitud võimalikult erinevat tüüpi uued ajalehed, mis iseloomustavad selle perioodi muutuvat keelekasutust kõige paremini: majandusleht Äripäev (asutatud 1989), arvamusleht Eesti Ekspress (1989), mis on olnud eesti ajalehenduse keskne muutja, maarahvale mõeldud Maaleht (1987), mis on suuruselt teine üldnädalaleht ning mille stiil ja lugejaskond erinevad kõige enam Ekspressi lugejaskonnast. Paljudest tabloidnädalalehtedest on valitud Liivimaa Kroonika, mis ilmus 1993–1998 ning seetõttu on esindatud ainult kahe alaperioodi valikus. Tabloidpäevalehed tekivad alles perioodi lõpul (Õhtuleht ja Sõnumileht) ning neid valikusse võetud ei ole.

3. Kohalike lehtede seast on uuena kahte viimasesse alaperioodi võetud kõige suurema tiraažiga kohalik leht Pärnu Postimees.

4. Nõukogude perioodi üleriigilised lehed jätkasid ilmumist kuni 1995. aastani. Seetõttu on nii Rahva Hääl kui Noorte Hääl/Päevaleht ka selle valiku kahte esimesse perioodi sisse võetud.

1995. aasta suvel ühinesid kolm üleriigilist päevalehte Rahva Hääl, Noorte Hääl ja Hommikuleht Eesti Päevaleheks. Sügisel neelas see alla ka Eesti Sõnumid, mille järel Sõnumite toimetus asutas uue päevalehe Sõnumileht, mis oli opositsioonis Päevalehega ning ainus vasakpoolne ajaleht Eestis kuni muutumiseni tabloidpäevaleheks 1998. Seetõttu on viimasel alaperioodil võetud korpusesse nii Eesti Päevaleht kui ka Sõnumileht enne tabloidistumist.

Tiraažid on perioodi algupoolelt teada väga juhuslikult ja kõiguvad palju aastate ning ka aastaaegade lõikes. Alates 1994. aastast on olemas korralikud ja pidevad andmed Eesti Ajalehtede Liidult (alates 1998. aastast kättesaadavad ka liidu koduleheküljel http://www.eall.ee).

Ajalehtede lõplik nimekiri tuli järgmine (nimestikus võib olla ka vigu, sest osa aastate tegelik valik ei ole praegu üle kontrollitud):

1890–1899: Olevik, Eesti Postimees, Postimees, Virmaline, Sakala, Ristirahva Pühapäeva-leht, Valgus

1900–1910: Olevik, Postimees, Teataja, Virulane, Sakala, Uus Aeg, Valgus, Saarlane

1911–1920: Vaba Maa, Postimees, Päevaleht, Sakala, Tallinna Teataja, Meie Aastasada, Kiir, Meie Elu, Olevik

1935–1939: Vaba Maa, Postimees, Päevaleht, Uudisleht, Uus Eesti, Rahvaleht, Maa Hääl, Esmaspäev, Ühistegel Uudised., Vaba Maa Pärnus, Järva Teataja, Oma Maa

1948–1952: Rahva Hääl, Noorte Hääl, Sirp ja Vasar, Punane Täht, Edasi, Õhtuleht, Talurahvaleht

1966–1970: Rahva Hääl, Noorte Hääl, Sirp ja Vasar, Punane Täht, Edasi, Õhtuleht, Kodumaa

1971–1975: Rahva Hääl, Noorte Hääl, Sirp ja Vasar, Punane Täht, Edasi, Õhtuleht, Kodumaa

1983–1987: Rahva Hääl, Noorte Hääl, Sirp ja Vasar, Punane Täht, Edasi, Õhtuleht, Kodumaa

1988–1991: Rahva Hääl, Noorte Hääl, Sirp ja Vasar, Punane Täht, Õhtuleht, Edasi/Postimees, Liivimaa Kroonika, Eesti Ekspress, Äripäev, Maaleht

1992–1995: Rahva Hääl, Päevaleht, Kultuurileht, Punane Täht, Õhtuleht, Postimees, Liivimaa Kroonika, Eesti Ekspress, Äripäev, Maaleht, Pärnu Postimees

1996–1998: Eesti Päevaleht, Sirp, Punane Täht, Õhtuleht, Postimees, Sõnumileht, Eesti Ekspress, Äripäev, Maaleht, Pärnu Postimees

2. Korpuse märgendus

Korpuse tegemise algusperioodil oli kavas kogu korpus märgendada eeskujukorpuse LOB märgenditega. Sellest loobuti peatselt ning mindi üle TEI märgendusele. TEI oli sel perioodil veel nõrgalt välja töötatud. Esialgne märgendus ei olnud üksüheselt võetud tollasest TEIst vaid oli selle põhjale tehtud omapoolne variatsioon. Alates 1995. aastast mindi üle TEI täpsele järgimisele. See nõudis osa varasema korpuse ümbermärgendamist. Praegu on kogu korpus sisestatud ja märgendatud vastavalt TEI reeglitele. Erinevad alakorpused on sealjuures märgendatud eri sügavusega.

2.1. TEI põhimõtted

TEI (*Text Encoding Initiative,* http://www.tei-c.org/) on rahvusvaheline uurimisprojekt, mille eesmärgiks on koostada ja levitada elektrooniliste tekstide ühtse märgendamise ja esitamise juhend (*Guidelines for Electronic Text Encoding and Interchange* (TEI-P3)).

TEI-s pole tõmmatud selget vahet objektiivse ja subjektiivse info või representatiivse ja interpretatiivse info vahel. Kuid osad märgendid esitavad selgelt teksti struktuuri (teksti osad, lõigud, laused) ja osad on interpretatiivsed, nt võimaldab TEI märgendada rõhutamise eesmärgil esiletõstetud tekstiosi märgendiga *<emph>*

TEI P3 on tegelikult üks SGMLi variante (SGML – *Standard General Markup Language*). SGMLi põhimõisteks on märgend, mis koosneb nurksulgudest ja kokkuleppelisest koodist (*<tag>*). Tavaliselt tähistatakse märgendite abil märgendatava elemendi algus ja lõpp. Kui soovitakse märgendatavat nähtust põhjalikumalt iseloomustada, siis lisatakse märgendile atribuudid. Nt märgendil *<name>* võib olla atribuut *<name type= >* ja atribuudil omakorda väärtused, nt *person, place.*

TEI soovitab igale korpuse tekstile ja ka kogu korpusele lisada päise *(Header),* mis identifitseerib, dokumenteerib ja kirjeldab korpuses olevaid tekste ning mille abil saab uurija valida korpusest talle vajalike omadustega tekste.

Väga vähesed märgendid TEI suurest märgendite hulgast on kohustuslikud selleks, et märgendatud tekst vastaks formaalselt TEI nõuetele. Enamus märgendeid ja nende atribuute on vabatahtlikud

so. neid kasutatakse vajadusel ja võimalusel. TEI märgendite hierarhia on järgmine:

- tekstiüksuste märgendid (*chunks*);
- lõigud ja teised lõigutasandi märgendid, mis võivad olla kas ainult teksti osad või teksti alljaotuste (<*div*>) osad, kuid ei või esineda teiste tekstiüksuste märgendite sees;
- märgendid, mis võivad esineda ainult lõigumärgendite või teiste lõigutasandi märgendite sees, mitte väljaspool neid (*phrase-level elements*);
- märgendid, mis võivad esineda kas lõikude vahel võrdselt lõigutasandi märgenditega või ka lõigutasandi märgendite sees (*inter-level elements*, nt loendi märgend <*list*>)

Mõned elemendid ei kuulu ühtegi eelpoolnimetatud klassi, nt sellised kogu teksti liigendamiseks kasutatavad märgendid nagu <*tei.2*> ja <*group*>.

Seega tuleb märgendamisel järgida märgendite teatud kindlat hierarhiat. Nt ei tohi paljud märgendid ulatuda üle lõigu <*p*> piiridest. Kui nt šrifti muutus <*hi rend*=...> kestab üle mitme lõigu, tuleb märgend enne lõiku lõpetavat märgendit </*p*> lõpetada ja uue lõigu alguses jälle alustada. Kogu teksti põhiosa peab olema tähistatud märgendiga <*body*> Praktiliselt tähendab see seda, et kõik automaatselt või käsitsi märgendatud tekstid tuleb spetsiaalse SGML-kontrollijaga üle kontrollida, enne kui võib öelda, et nad on formaalselt korrektsed.

Tekstis endas soovitab TEI märgendada näiteks järgmisi nähtusi:

- lõigust suuremad alajaotused <*div*>, lõigud <*p*>, laused <*s*>, osalause <*cl*>, fraas <*phr*>;
- pealkirjad <*head*>, teksti allikas <*bibl*>, autor <*author*>, loendid <*list*> tabelid <*table*>;
- lühendid <*abbr*>;
- tsitaadid ja muu jutumärkides olev materjal <*q*>, <*quote*>, <*cit*>, <*soCalled*>;
- pärisnimed <*name*>;
- võõrkeelsed sõnad ja väljendid <*foreign*>;
- kuupäevad, aastaarvud <*date*>, kellajad <*time*>;
- väljajäetav materjal: illustratsioonid, tabelid jne <*gap*>.

2.2. Internetis olevate korpuste märgendus

Baaskorpus on Internetis olemas kolmes versioonis.

1. Märgendamata tekst, iga lause eraldi real so. samal kujul nagu ülejäänud aastakümned.

2. TEI järgi märgendatud tekst, mille poolest jagunevad tekstiklassid kahte rühma.

Esimese rühma moodustavad ajakirjandustekstid, ilukirjandustekstid ja teadustekstid. Neis on märgendatud nii vormilisi (pealkirjad, allkirjad, teksti osad, lõigud, laused, šrifti muutused, tabelid jms) kui ka sisulisi nähtusi (lühendid + nende tähendus, pärisnimed + nende liik, arvud, kuupäevad, kellaajad, mõõtühikud, jutumärgid ja nende otstarve jms.) Seega on märgendatud võimalikult paljusid nähtusi, mis võivad tekstis käituda teistmoodi kui tavalised tekstisõnad. See märgendus on tehtud suures osas käsitsi ja on küll formaalselt korrektne, kuid sisuliselt veidi ebaühtlane.

Teise rühma moodustavad baaskorpuse ülejäänud tekstiklassid (Populaarteadus, Esseed ja biograafiad, Hobi- ja harrastustekstid, Propaganda, Entsüklopeedilised tekstid, Dokumendid ja Vaimulikud tekstid). Nendes on loobutud enamuse sisuliste nähtuste (pärisnimed, lühendid, kuupäevad, kellaajad, arvud, mõõtühikud) märgendamisest ja on märgendatud ainult teksti struktuuri (teksti osad, pealkirjad, allkirjad, lõigud, laused, šrifti muutused, tabelid jms). Põhjuseks on eeskätt see, et põhjalik märgendamine nõudis väga palju aega, tööjõudu ja raha ja teiseks tuleks sellisel märgendamisel kindlasti varasemast enam pöörata tähelepanu märgendamise ühtlustamisele.

Kõikidele tekstidele on lisatud päis (*Header*), kus on kirjas info selle faili allika (paberkandjal oleva teksti) kohta ja failis tehtud muutuste (põhiliselt märgendamise ja märgendajate) kohta.

3. Morfoloogiliselt analüüsitud (aga mitte ühestatud) versioon, mille tegi Leho Paldre.

Kõik varasemad korpused on märgendatud ühte moodi: lause real, pealkirjad ja allkirjad nagu laused so. lause tasandile. 1990. aastate TEI korpus on märgendatud samamoodi nagu varasemad korpused. ELANi tekstides on märgendatud TEI järgi peatükid, pealkirjad, allkirjad, lõigud, laused, tabelid ja šrifti muutused.

Lausestamine on tehtud automaatselt ja inimese poolt on seda üle kontrollitud ainult baaskorpuse puhul, seega võib lausestamises olla vigu.

3. Kaks allkeelt

Järgnevalt vaatleme lühidalt kahe allkeele sagedaste sõnade sageduste muutumisi 20. sajandi tekstides. Uuritavaks on võetud kuus allkorpust: 1930., 1960. ja 1990. aastate ajakirjandus ja ilukirjandus. 1990. aastatest on võrreldavuse huvides vaadeldud Tiit Hennoste valikute järgi koostatud korpust, millle tekstivaliku kriteeriumid on samad kui varasematel korpustel. Kõrvale on jäetud ELANi ajakirjandustekstid, mille valikukriteeriumid on teistsugused. Korpuste suurused on järgmised:

- ajakirjandus: 1930. aastad: 117 000; 1960. aastad: 201 000; 1990. aastad: 384 800;
- ilukirjandus: 1930. aastad: 252 000; 1960. aastad: 132 000; 1990. aastad: 611 000.

3.1. Statistika tegemise põhimõtted

Allkorpustest on tehtud kõigepealt sõna algvormide (lemmade) sagedussõnastikud, millest on valitud võrdluseks 100 kõige sagedasemat lemmat (vt lemmade tabelit Lisas).

Valitud alamkorpused töödeldi mõnede Unixi skriptide abil morfanalüsaatorile sobivale kujule. Seejärel lisati eesti keele morfanalüsaatori ESTMORF abil igale sõnavormile kõik selle võimalikud morfoloogilised tõlgendused. ESTMORF töötas nn oletamisrežiimis, mis tähendab, et kõik sõnavormid, ka analüsaatorile tundmatud, said mingi tõlgenduse. Saadud morfoloogiliste analüüside hulgast antud kontekstis õige väljavalimiseks on kasutatud statistikale tuginevat eesti keele morfoloogilist ühestajat (morfanalüsaatori ja ühestaja kohta vt Kaalep 1998; Kaalep, Vaino 2000). Ühestaja teeb vigu 1,47%. Mitteühesed vormid ja vead kokku moodustavad 4,02% väljundist. Kuna praegu huvitasid meid ainult lemmade sagedused, siis võeti Unixi skriptide abil ühestaja väljundist välja ainult lemmad. Seejärel arvutati, millise osa moodustab iga lemma antud tekstiklassi lemmade üldarvust, mida väljendati promillides (‰).

3.2. Oluliste sõnade esinemissageduste võrdlusi

Järgnevalt vaatleme selliseid sõnu, mille sagedused esimese saja sõna seas kas erinevad ajakirjanduse ja ilukirjanduse alamkorpustes või mille sagedused on neis korpustes aja jooksul kindla trendiga muutunud.

Verbid. Kõige sagedasem verb on *olema*, mis on järgnevatest verbidest umbes 10 korda sagedasem (1990. ajakirjanduses nt 42,17‰, järgmine on *saama* 4,90‰).

Ajakirjanduses on sageduse ülemises otsas modaalverbid (*saama, pidama, tulema, võima*). Neile lisanduvad *jääma* ja *hakkama* ilmselt kui fraasiverbid, väga üldise sisuga verbid *tegema* ja *minema*, keskne otsese kõne saatelause verb verb *ütlema* ja ka verbid *võtma* ja *andma*, mille suurt sagedust on raske seletada. 1930. ja 1990. on sagedaste seas ka *tahtma*, *teadma* ja *arvama*, mida 1960. aastatel pole.

Ilukirjanduses on sageduse ülemises otsas modaalverbid (*saama, pidama, tulema*), samuti *ütlema, tulema, tegema* ja *minema*. Lisaks veel meeltetegevust väljendavad verbid *nägema, vaatama, mõtlema, tundma*. Seega erinevad ilukirjanduse ja ajakirjanduse sagedaste verbide tuumikud üksteisest eeskätt meeltetegevust väljendavate verbide kasutuse poolest.

Ajakirjanduses on verbe saja sagedasema lemma hulgas tunduvalt vähem kui ilukirjanduses. 1930. aastate ajakirjanduses on 100 sagedasema lemma hulgas verbe 16, 1960. – 11, 1990. – 15, ilukirjanduses vastavalt 22, 25 ja 24 verbi. Samuti on ajakirjanduses palju vähem eriti sagedasi verbe. Nt üle 1,5‰ sagedusega verbe on ajakirjanduses 9–12 ja ilukirjanduses 17–21. Seega on ajakirjandus kõigil perioodidel oluliselt nominatiivsem, ehk lähemal asjaajamise ja teaduslikule allkeelele. Samasugune pilt avaneb ka teiste keelte materjalides.

Nimisõnad. Sagedaste nimisõnde poolest erinevad ajakirjandus ja ilukirjandus üksteisest tunduvalt. Ilukirjanduses on kõige sagedasem sõna *mees* (3,16 → 3,99 → 3,55) [1] mille sagedus on ajakirjanduses

[1] Sagedust osutavates ridades on alati esikohal 1930. aastate korpuse sagedus promillides, teisel kohal 1960. aastate korpuse sagedus ja kolmandal 1990. aastate korpuse sagedus.

üle kahe korra madalam (1,61 → 0,78 (136. kohal) → 1,22). Lisaks kuuluvad sagedusrea ülemisse otsa (sagedusega üle 1,5‰) peaaegu ainult mõned teised inimesega seotud sõnad: *käsi, inimene, naine, silm, pea* ja 1960. aastatel ka sõnad *aeg* ja *päev*. *aeg* (ajakirjanduses: 2,43 → 2,25 → 2,44; ilukirjanduses: 2,48 → 2,83 → 2,64) on sealjuures kasutatud põhiliselt fraasistunud väljendites nagu *sel ajal, viimasel ajal*. Lisaks on ilukirjanduse nimisõnade loend ja sagedused kõigil kolmel perioodil väga sarnased. Ainus oluline muutus on sõna *naine* sageduse tõus 1990. aastatel (2,04 → 2,05 → 2,32).

Ajakirjanduse nimisõnade loendite algused erinevad ilukirjanduse omadest ja jagunevad kahte poliitiliselt määratud rühma. 1930. ja 1990. aastatel on sagedusrea alguses *aasta, aeg, Eesti, Tallinn, kroon, mees, päev, inimene*. Üle 1,5‰ olevaid verbe on neis loendites 8–10. 1960. aastate algusots jaguneb aga kahte rühma. Ühte kuuluvad osalt needsamad sõnad *aasta, Eesti, aeg, inimene, Tallinn* ja sisuliselt samasse rühma kuuluvad kohanimetused *NSV, Liit, rajoon*, teise aga kommunistliku ideoloogia ja riigiehituse olulised mõisted (*töö, nõukogu, partei, rahvas, kommunistlik, keskkomitee, maa, kolhoos, töötaja*). Üle 1,5‰ olevaid verbe on sel perioodil 18. Eriti huvitav on see, et erinevused kahe perioodirühma ühiste verbide sagedustes on väikesed. Nõukogulik sõnavara tuleb otsekui lisaks, ilma et ta asendaks muid keskseid nimisõnu.

1960. aastate ajakirjanduses oli 100 sagedasema lemma hulgas kõige vähem verbe (11) ja kõige rohkem nimisõnu (37), ehk selle perioodi ajakirjanduskeel on kõige nominatiivsem. Nominatiivsus iseloomustab tavaliselt ametlikku, bürokraatlikku või teaduslikku keelekasutust ja seda heideti ajakirjanduskeelele nõukogude ajal pidevalt ette. Samal ajal aga on selle perioodi ilukirjanduses verbide esinemine ja sagedus suurem kui 1930. ja 1990. aastatel. Teisisõnu, selle perioodi ajakirjandus on kõige nominatiivsem ja ilukirjandus kõige verbikesksem. Muude perioodide allkeeled on üksteisele lähedasemad.

Asesõnad. Asesõnade kasutust iseloomustab eelkõige see, et neid on ilukirjandustekstides oluliselt enam kui ajakirjanduses (asesõnade liigutuse kohta vt EKG I 1995).

Isikulised asesõnad

mina (ajakirjanduses: 5,82 → 7,7 → 8,94, ilukirjanduses: 15,14 → 21,19 → 21,70) tõuseb mõlemas allkeeles. Ajakirjanduses on tõus ühtlasem, ilukirjanduses toimub suur hüpe 1930. ja 1960. vahel (6‰). Sealjuures kasutatakse *mina* ilukirjanduses pidevalt 10–15‰ rohkem kui ajakirjanduses. Ilukirjanduses on *mina* kasutushüpe seletatav eelkõige *mina*-vormis ilukirjandusteoste (*mina*-jutustuse) hulga suure kasvuga just 1960. aastatel. Ajakirjanduses on pidev tõus raskemini seletatav. Selle taga saab olla dialoogi (tsitaatide) hulga kasv ning arvamuslugude ja olemuslugude hulga kasv ajakirjanduses.

sina (ajakirjanduses: 1,55 → 0,86 (pole 100 hulgas) → 1,25, ilukirjanduses: 9,56 → 10,43 → 9,01) muutub ajakirjanduses ja ilukirjanduses erinevalt: ajakirjanduses liigub 1960. alla, ilukirjanduses üles. Aga mõlemas tekstiklassis on seda 1990. aastatel vähem kui 1930. Ilukirjanduses kasutatakse *sina* 8–9‰ enam kui ajakirjanduses, mis on loomulik, sest *sina* on dialoogisõna, mida ajakirjanduses on väga raske kasutada.

tema ja *see* on võetud kokku, sest nende mitmuse vorme on statistilisel ühestajal võimatu lemmatiseerida.

tema (ajakirjanduses: 11,71 → 9,00 → 8,25), *see* (ajakirjanduses: 19,90 → 18,40 → 19,80)

tema (ilukirjanduses: 37,09 → 28,44 → 29,84), *see* (ilukirjanduses: 27,40 → 23,41 → 24,94)

Ka neid sõnu on ilukirjanduses palju enam kui ajakirjanduses. Sealjuures on *tema* ilukirjanduses üle 20‰ enam, *see* üle 5‰ enam. Ajakirjanduses väheneb *tema* kokku umbes 3‰ võrra, *see* muutub väga vähe. Ilukirjanduses langevad nii *see* kui *tema* 1960. aastatel 4–9‰. See on selges korrelatsioonis *mina* kasutussageduse muutumisega. Ajakirjanduses tõuseb *mina* kasutus pidevalt, 1990. aastatel on ta isegi *tema*st möödunud.

Omastav asesõna *oma* (ajakirjanduses: 5,82 → 5,03 → 5,32, ilukirjanduses: 8,36 → 6,05 → 6,74) langeb tugevalt peale 1930. aastaid. Miks, vajab eraldi uurimist.

Määratlevad asesõnad

ise (ajakirjanduses: 2,65 → 1,71 → 2,93, ilukirjanduses: 6,18 → 5,67 → 5,7).

kõik (ajakirjanduses: 2,80 → 3,63 → 3,05, ilukirjanduses: 3,50 → 3,64 → 3,86)

kogu (ajakirjanduses: 1,17 → 1,62 → 1,00, ilukirjanduses: 1,19 → 0,90 (pole 100 hulgas) →0,93 (pole 100 hulgas))

Määratlevad asesõnad käituvad mitmesuguselt, ühest loogikat nende muutustes ei ole.

Umbmäärased asesõnad

miski (ajakirjanduses: 0,70 → 0,56 → 1,12, ilukirjanduses: 3,10 → 3,15 → 3,44). Ajakirjanduses pole 1930. ja 1960. 100 sagedasema hulgas.

keegi (ajakirjanduses: 0,66 → 0,30 → 0,77 (pole 100 sagedasema hulgas), ilukirjanduses: 1,36 → 1,41 → 1,89).

üks (ajakirjanduses: 3,16 → 2,14 → 3,00, ilukirjanduses: 2,79 → 3,12 → 3,39).

teine (ajakirjanduses: 3,10 → 3,02 → 2,49, ilukirjanduses: 3,10 → 2,64 → 2,68).

mõni (ajakirjanduses: 1,54 → 1,13 → 1,35, ilukirjanduses: 1,90 → 1,74 → 1,88).

Korpuse käsitsi morfoloogilise ühestamise põhjal võib väita, et lemmasid *üks* ja *teine* kasutatakse põhiliselt umbmääraste asesõnadena (vt Kaalep, Muischnek, Müürisep, Rääbis, Habicht 2000). Tuldava jt järgi on aga 1960. aastate ilukirjanduse autoritekstis 304 *teine* esinemisest 185 asesõnalised ja 119 arvsõnalised, *üks* 350 esinemisest 252 arvsõnalised, 94 asesõnalised ja 4 numeraali käändelist vormi väljendverbi komponendina (Tuldava jt 1977). Erinevused võivad siin olla põhjustatud sellest, et Tuldava kasutas ainult autorikõne korpust, kuid nende sõnade kasutamine asesõnadena on levinud just dialoogis. Teiseks võib oletada, et neid sõnu kasutatakse ilukirjanduses ja ajakirjanduses erinevalt. Ilukirjanduses võib oodata suuremat umbmääraste asesõnade kasutust, ajakirjanduses aga võib eeldada, et neid kasutatakse enam numbritena.

Umbmäärased asesõnad (peale sõna *teine*) tõusevad ilukirjanduses pidevalt, eriti 1990. aastatel. Kuna umbmäärased asesõnad iseloomustavad suulist kõnet ja argikeelt, siis võib selle taustaks olla

argikeelsuse ja dialoogi osa oluline suurenemine 1990. aastate ilukirjanduses. Ajakirjanduses väheneb nende kasutus 1960. aastatel.

Näitavatest asesõnadest on 100 sagedasema lemma hulgas pidevalt vaid *see* ja *tema* (vt eespool). Vastastikuseid ja enesekohaseid asesõnu pole 100 sagedasema lemma hulgas. Küsivad–siduvad asesõnad tulevad jutuks sidendite osas.

Asesõnu on ilukirjanduses selgelt enam kui ajakirjanduses, eriti isikulisi asesõnu. Selle taustaks on ilukirjanduse suurem dialoogilisus ja ka suurem lähedus suulisele kõnele (vt ka Hennoste jt 2000).

Muutumatud sõnad

Ajaadverbid

siis (ajakirjanduses: 3,49 → 1,77 → 2,97, ilukirjanduses: 8,25 → 6,29 → 5,98). *kui...siis* konstruktsioone on *siis* kasutusest umbes 21–29% eri aegade korpustes (võimalik arvutusviga on suur).

nüüd (ajakirjanduses: 1,92 → 1,00 → 1,31, ilukirjanduses: 5,10 → 3,37 → 2,69).

siis ja *nüüd* käituvad ühtmoodi: langevad tugevalt peale 1930. aastaid. Ajakirjanduses toimub sealjuures 1990. väike tõus, ilukirjanduses jätkub langus. Sellise käitumise põhjused vajavad veel uurimist.

pärast (ajakirjanduses: 1,21 → 1,14 → 1,25, ilukirjanduses: 1,45 → 2,38 → 1,54).

Sellel sõnal on 2 funktsiooni – prepositsiooni ja adverbina väljendab ta aega, postpositsioonina põhjust vms. Tuldava andmetel (Tuldava jt 1977) on 190 sõnast *pärast* 166 adpositsioonid ja 24 iseseisvad adverbid. Kuna käesolevas on muutumatud sõnad ühte klassi võetud, siis ei tehta siin vahet kaassõna ja adverbi vahel.

praegu (ajakirjanduses: 1,46 → 1,19 → 1,24, ilukirjanduses pole 100 sagedasema lemma hulgas)

juba (ajakirjanduses: 2,79 → 2,01 → 1,87, ilukirjanduses: 3,52 → 3,09 → 2,63)

Kohaadverbid. Kõik kohaadverbid va. *seal* ja *siin* esinevad ka ühendverbi komponendina ja kaassõnana. Iseseisva adverbina ja kaassõnana on neil kohatähendus, ühendverbi komponendina võib ka mitte olla. Praeguse meetodiga ei saa neid eristada. 1960. aastate ilukirjanduse autorikõnes kasutatakse neid sõnu valdavalt adpositsioonina või ühendverbi osana, mitte iseseisva adverbina (Tuldava jt 1977).

siin (ajakirjanduses: 0,92 (102. kohal) → 0,96 (104. kohal) → 0,92 (107. kohal), ilukirjanduses: 2,44 → 1,93 → 1,80)

seal (ajakirjanduses: 1,07 → 0,5 (239. kohal) → 0,72 (140. kohal), ilukirjanduses: 1,60 → 1,61 → 1,53)

siin ja *seal* esinevad ilukirjanduses selgelt sagedamini kui ajakirjanduses. Nende kasutus on laias laastus kogu aeg vähenenud.

Adverbid *siis, nüüd, siin* ja *seal* on väga laia ja üldise tähendusega verbid, mis iseloomustavad suulist kõnet. Nende pidev vähenemine iseloomustab seega kirjaliku keele pidevat kaugenemist suulisest kõnest selles punktis (vt võrdluseks Hennoste jt 2000).

välja (ajakirjanduses: 1,68 → 1,51 → 1,77. ilukirjanduses: 1,67 → 2,21 → 2,29). Tuldaval jt on *välja* kasutusest umbes 90% ühendverbi osa.

üle (ajakirjanduses: 1,98 → 1,81 → 1,62, ilukirjanduses: 2,11 → 2,02 → 1,88). Tuldaval jt 78% adpositsioonid, 21 % iseseisvad adverbid.

vastu (ajakirjanduses: 2,08 → 2,31 → 1,44, ilukirjanduses: 2,09 → 1,81 → 1,69). Tuldaval jt 61% adpositsioonid ja muud iseseisvad määrsõnad.

eest (ajakirjanduses: 1,35 → 1,90 → 1,38, ilukirjanduses 100. hulgast väljas). Tuldaval jt 93% adpositsioonid.

peale (ajakirjanduses: 1,34 → 111. kohal → 116. kohal, ilukirjanduses 100 sagedasema hulgast väljas). Tuldaval jt 65% adpositsioonid, 34% ühendverbi osad.

ette (ajakirjanduses: 1,16 →135. kohal → 199. kohal, ilukirjanduses 100 sagedasema hulgast väljas). Tuldaval jt 56% ühendverbi osad ja 38% adpositsioonid.

tagasi (ajakirjanduses: 1,04 →0,65 (172. kohal) → 0,96, ilukirjanduses: 1,32 → 1,63 → 1,49). Tuldaval jt 75% ühendverbi osa, 20% adpositsioonid.

Ühendaverbides sisalduvaid adverbe on ilukirjanduses ja ajakirjanduses suhteliselt ühepalju. Muutused nende adverbide kasutussageduses on mitmesugused, kuid üldiselt väikesed. Ka pole praegu selge, kas ja kuidas muutub nende iseseisvana ja adpositsioonina või ühendverbi osana kasutamise suhe.

Sidendid (lisaks sidesõnadele on siia võetud ka siduvad asesõnad ja määrsõnad). Sidendeid on ilukirjanduses üldiselt enam kui ajakirjanduses, ajakirjanduses on enam vaid *kuid* ja *kuna*. Suhteliselt võrdselt on sõnu *ning, vaid, et* ja *kes*.

Rinnastavad sidesõnad

a) üldühendavad sidesõnad

ja (ajakirjanduses: 29,71 → 34,27 → 28,28, ilukirjanduses: 36,44 → 31,69 → 30,23)

ning (ajakirjanduses: 3,25 → 4,24 → 4,52, ilukirjanduses: 4,46 → 2,74 → 3,95)

Need sõnad on sünonüümid, millest *ning* kasutus on just kirjakeelele omane nähtus (vt Hennoste jt 2000). Samas käituvad need sõnad 1990. aastatel erinevalt kui varasematel perioodidel. 1990. aastatel võib neid vaadata kui üksteist asendavaid sõnu (*ja* langeb ja *ning* tõuseb), kuid varasematel perioodidel sellist asendust ei esine. *ja* ja *ning* kasutus kokku tõuseb tugevalt 1960. aastate ajakirjanduses ja langeb 1960. ja 1990. aastate ilukirjanduses võrreldes 1930. aastatega. Sealjuures kasutatakse neid ilukirjanduses selgelt enam kui ajakirjanduses.

b) vastandavad sidesõnad

aga (ajakirjanduses: 3,83 → 2,99 → 4,49, ilukirjanduses: 7,65 → 7,85 → 7,12)

kuid (ajakirjanduses: 2,38 → 2,00 → 1,79 , ilukirjanduses: 3,16 → 2,05 → 1,97)

vaid (ajakirjanduses: 1,67 → 1,08 → 1,71, ilukirjanduses: 1,67 → 139. kohal (0,89) → 1,57)

c) eraldav sidesõna *või* (ajakirjanduses: 2,29 → 1,63 → 3,27, ilukirjanduses: 3,04 → 3,57 → 3,80)

Vastandavad ja eraldavad sidesõnad käituvad suhteliselt ühtemoodi. Nende kasutusagedus langeb 1960. aastate ajakirjanduses ja tõuseb taas 1990. aastateks (v.a *kuid*). Ilukirjanduses käituvad nad mitmel viisil, milles on raske mingit ühtset loogikat leida.

Alistavad sidesõnad

et (ajakirjanduses: 11,31 → 9,53 → 11,35, ilukirjanduses: 12,21 → 12,77 → 12,34)

et-kõrvallaused on peamiselt komplementlaused, mis laiendavad harilikult psühho- või suhtlussõnu. Siinkohal tuleks mainida, et 1990. aastate ajakirjanduses on 100 sagedasema lemma hulgas verbid *arvama, teadma, tahtma*; 1930. ajakirjanduses *tahtma, leidma, nägema, arvama,* millest ühtegi ei ole 1960. ajakirjanduse 100 sagedasema lemma hulgas.

sest (ajakirjanduses: 1,26 → 0,73 (148. koht) → 1,20, ilukirjanduses: 1,91 → 1,52 → 1,80)

nagu (ajakirjanduses: 1,91 → 1,84 → 1,63, ilukirjanduses: 5,95 → 5,05 → 4,61)

Alistav-rinnastavad sidesõnad

kui (ajakirjanduses: 7,27 → 4,68 → 7,39, ilukirjanduses: 10,79 → 10,78 → 9,55).

ei (ajakirjanduses: 8,79 → 6,47 → 10,04, ilukirjanduses: 15,17 → 16,16 → 16,95)

Lisaks sidesõna funktsioonile osaleb *ei* ka verbi eitava vormi moodustamisel. Tuldava järgi on *ei* 1395 esinemisest 1330 rõhumäärsõna (95,3%), 45 interjektsiooni ja 20 ühendsidesõna komponenti. Ka selles korpuses on rõhuv enamus eitav kõne.

kus (ajakirjanduses: 1,71 → 1,34 → 1,71 , ilukirjanduses: 1,41 → 1,34 → 1,47)

kes (ajakirjanduses: 5,30 → 4,00 → 4,33, ilukirjanduses: 4,94 → 4,42 → 4,61)

mis (ajakirjanduses: 8,70 → 7,50 → 7,58, ilukirjanduses: 9,84 → 8,70 → 9,20)

Alistavad, alistav-rinnastavad, vastandavad ja eraldavad sidesõnad käituvad ajakirjanduses ühtmoodi (va. *nagu*). Kõigi nende kasutus väheneb 1960. aastatel, et peale seda taas tõusta. Need sõnad annavad edasi vastandusi, põhjusi ja tingimusi, seega võib nende kasutuse vähenemine olla korrelatsioonis sellega, et nõukogude perioodi ajakirjandus tegeles vähem kui muudel aegadel põhjuste ja tingimuste väljatoomisega ning vastandamisega. See on usutav, sest on hästi sobiv nõukogude ideoloogia eesmärkidega. Ilukirjanduses muutuvad need sõnad mitmel viisil. Kaks põhirühma on vähenemine (*mis, kes, kus, kuna, sest*) ja püsimine (*aga, et, sest, ei*). 1960. aastate ajakirjanduses langevad sageduselt ka ajamäärsõnad *siis, nüüd, enam, juba*; asesõnad *sina, tema, see, oma, üks*. Tõusevad asesõna *mina* ja sidesõnad *ja* ja *ning*. Mis on selle põhjuseks, vajab eraldi uurimist.

Kirjandus

Atkins S., Clear, J. and Ostler, N. 1992. Corpus design criteria. – Literary and Linguistic Computing, 7 (1), 1–16.

Biber, D. A. 1989. Typology of English texts. – Linguistics, 27 (1), 3–43.

EKG I 1995 = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. Eesti keele grammatika. I. Morfoloogia. Sõnamoodustus. Tallinn: ETA Eesti Keele Instituut.

Hennoste, T. 1996. Tartu University Corpus of Written Estonian: A survey of the structure of texts and principles of selection. – Estonian in the Changing World. University of Tartu. Toim H. Õim. Department of General Linguistics. Tartu. 7–32.

Hennoste, T., Koit, M., Roosmaa, T., Saluveer, M. 1998. Structure and usage of the Tartu University Corpus of Written Estonian. – International Journal of Corpus Linguistics 3 (2), 279–304.

Hennoste, T., Muischnek, K., Potter, H., Roosmaa, T. 1993. Tartu Ülikooli eesti kirjakeele korpus: ülevaade tehtust ja probleemidest. – Keel ja Kirjandus 10, 587–600.

Hennoste, T., Lindström, L. Rääbis, A., Toomet, P., Vellerind, R. 2000. Eesti suulise kõne korpus ja mõnede allkeelte võrdluse katse. – Käesolevas kogumikus.

Kaalep, H.-J. 1998. Tekstikorpuse abil loodud eesti keele morfoloogia-analüsaator. – Keel ja Kirjandus 1, 22–29.

Kaalep, H.-J., Vaino, T. 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Käesolevas kogumikus.

Kaalep, H.-J., Muischnek, K., Müürisep, K., Rääbis, A., Habicht, K. 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? – Keel ja Kirjandus, ilmumas.

Kruus, O. (koost) 1995. Eesti kirjarahva leksikon. Tallinn: Eesti Raamat.

Lauk, E. 1991. Eesti Vabariigi ajakirjandusest 1920.–1930.aastatel. – Eesti ajakirjanduse ajaloost VII. Tartu: Tartu Ülikool. 36–78.

McEnery, T., Wilson, A. 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.

Peegel, J., Aru, K., Issakov, S., Jansen, E., Lauk, E. 1994. Eesti ajakirjanduse teed ja ristteed. Eesti ajakirjanduse arengust (XVII sajandist XX sajandini). Tallinn: Olion.

Sinclair, J. M. 1991. Corpus, Concordance, Collocation. Describing English Language. Oxford: Oxford UP.

Sperberg-McQueen, M., Burnard. L. (toim). 1994. Text Encoding Initiative. Guidelines for Electronic Text Encoding and Interchange. Chicago, Oxford.

Tuldava, J., Kaasik, Ü., Villup, A., Ääremaa, K. 1977. Eesti tänapäeva ilukirjandusproosa autorikõne lekseemide sagedussõnastik. – Töid keelestatistika alalt II (Tartu Riikliku Ülikooli toimetised. Vihik 413) Tartu: Tartu Riiklik Ülikool. 5–140.

Lisa. Ajakirjandus- ja ilukirjandustekstide lemmade sagedused 1930., 1960. ja 1990. aastatel. 100 sagedasimat lemmat

Ilukirjandustekstide sõnad								
1930			1960			1990		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
olema	10966	43,52	olema	5764	43,67	olema	28237	46,21
tema	9347	37,09	ja	4183	31,69	ja	18469	30,23
ja	9184	36,44	tema	3754	28,44	tema	18235	29,84
see	6904	27,40	see	3090	23,41	see	15240	24,94
el	3822	15,17	mina	2797	21,19	mina	13261	21,70
mina	3816	15,14	el	2133	16,16	el	10360	16,95
et	3077	12,21	et	1685	12,77	et	7538	12,34
kui	2720	10,79	kui	1423	10,78	kui	5833	9,55
mis	2480	9,84	sina	1377	10,43	mis	5621	9,20
sina	2408	9,56	mis	1149	8,71	sina	5505	9,01
oma	2106	8,36	aga	1036	7,50	aga	4350	7,12
siis	2080	8,25	siis	830	6,29	oma	4120	6,74
aga	1927	7,65	oma	799	6,05	siis	3655	5,98
ise	1556	6,18	ise	749	5,67	ise	3403	5,70
nagu	1500	5,95	nagu	666	5,05	ka	2992	4,90
nüüd	1285	5,10	tulema	652	4,94	kes	2819	4,61
kes	1245	4,94	minema	645	4,89	nagu	2814	4,61
nii	1207	4,79	ütlema	641	4,86	nii	2589	4,24
ning	1124	4,46	ka	620	4,70	tulema	2401	4,08
ütlema	1116	4,43	kes	583	4,42	saama	2474	4,05
tulema	1085	4,31	nii	568	4,30	ütlema	2419	3,96
minema	1064	4,22	saama	567	4,30	ning	2416	3,95
ka	1037	4,12	mees	526	3,99	minema	2378	3,89
veel	976	3,87	pidama	508	3,85	kõik	2360	3,86
pidama	893	3,54	veel	496	3,76	või	2321	3,80
juba	886	3,52	kõik	480	3,64	mees	2169	3,55
kõik	883	3,50	või	472	3,58	pidama	2122	3,47
saama	830	3,29	nüüd	445	3,37	miski	2101	3,44
mees	797	3,16	tegema	420	3,18	veel	2096	3,43
kuid	793	3,16	miski	416	3,15	üks	2072	3,39
teine	782	3,10	hakkama	413	3,13	tegema	1933	3,15
miski	781	3,10	üks	412	3,12	kas	1839	3,01
hakkama	772	3,06	teadma	411	3,12	teadma	1677	2,74
või	766	3,04	juba	408	3,09	nüüd	1641	2,69
tegema	743	2,95	kas	399	3,02	teine	1637	2,68

Ilukirjandustekstide sõnad								
1930			1960			1990		
Sõna	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**
kas	734	2,91	võtma	373	2,83	aeg	1610	2,64
nägema	724	2,87	aeg	373	2,83	juba	1605	2,63
üks	702	2,79	käsi	369	2,80	hakkama	1499	2,45
käsi	701	2,78	inimene	363	2,75	võima	1480	2,42
aeg	624	2,48	ning	362	2,74	vaatama	1477	2,42
siin	614	2,44	vaatama	361	2,73	võtma	1440	2,36
enam	613	2,43	teine	348	2,64	naine	1420	2,32
teadma	596	2,37	küll	339	2,57	käsi	1419	2,32
võima	593	2,35	ainult	327	2,48	välja	1400	2,29
vaatama	588	2,33	jääma	320	2,42	jääma	1380	2,26
suur	577	2,29	võima	315	2,39	küll	1334	2,18
tahtma	569	2,26	pärast	314	2,38	nägema	1329	2,18
küll	553	2,19	tahtma	297	2,25	inimene	1298	2,12
ju	548	2,17	ära	294	2,23	mitte	1283	2,10
inimene	545	2,16	nägema	294	2,23	tahtma	1238	2,07
ikka	534	2,12	välja	292	2,21	ära	1229	2,01
üle	532	2,11	enam	288	2,18	kuid	1201	1,97
võtma	528	2,10	ju	275	2,08	keegi	1157	1,89
vastu	527	2,09	naine	271	2,05	enam	1151	1,88
ega	519	2,06	kuid	271	2,05	üle	1150	1,88
jääma	517	2,05	mõtlema	270	2,05	mõni	1148	1,88
naine	514	2,04	üle	267	2,02	ju	1145	1,87
kord	502	1,99	pea	267	2,02	kord	1126	1,84
mitte	500	1,98	rääkima	266	2,02	suur	1118	1,83
olnud	482	1,91	peale	261	1,98	siin	1097	1,80
sest	481	1,91	kord	260	1,97	sest	1097	1,80
mõni	479	1,90	siin	255	1,93	olnud	1095	1,79
jälle	472	1,87	ega	255	1,93	ainult	1082	1,78
mõtlema	467	1,85	ikka	245	1,86	pea	1062	1,74
silm	452	1,79	vastu	239	1,81	ega	1046	1,71
tundma	445	1,77	andma	234	1,77	vastu	1030	1,69
ainult	440	1,75	mõni	230	1,74	kuidas	996	1,63
välja	422	1,67	asi	230	1,74	silm	990	1,62
vaid	421	1,67	küsima	228	1,73	mõtlema	982	1,61
kuidas	418	1,66	suur	223	1,69	vaid	960	1,57
asi	410	1,63	mitte	221	1,67	peale	956	1,56
seal	404	1,60	silm	218	1,65	rääkima	946	1,55
pea	390	1,55	ema	216	1,64	pärast	943	1,54
palju	388	1,54	tagasi	215	1,63	seal	935	1,53
nägu	370	1,47	käima	215	1,63	tagasi	910	1,49
pärast	365	1,45	seal	213	1,61	ikka	909	1,49
vana	361	1,43	hea	213	1,61	kus	901	1,47
kus	356	1,41	panema	209	1,58	pool	900	1,47
seisma	355	1,41	palju	204	1,55	asi	899	1,47
elu	353	1,40	päev	203	1,54	tundma	883	1,45
keegi	342	1,36	sest	200	1,52	andma	876	1,43
läbi	341	1,35	kuidas	198	1,50	päev	872	1,43
istuma	334	1,33	aasta	195	1,48	elu	859	1,41

Ilukirjandustekstide sõnad								
1930			1960			1990		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
tagasi	332	1,32	seisma	188	1,42	panema	856	1,40
sõna	331	1,31	keegi	187	1,41	küsima	855	1,40
isa	330	1,31	pool	180	1,36	mingi	848	1,39
kodu	328	1,30	iga	179	1,36	palju	844	1,38
küsima	326	1,29	isa	178	1,35	seisma	839	1,37
iga	321	1,27	kus	177	1,34	aasta	834	1,36
kohe	315	1,25	istuma	175	1,33	kaks	824	1,35
päev	314	1,25	töö	172	1,30	ema	814	1,33
andma	314	1,25	nagu	170	1,28	läbi	781	1,28
jah	312	1,24	jälle	168	1,27	iga	766	1,25
pool	304	1,21	läbi	166	1,26	käima	753	1,23
ema	301	1,19	tundma	165	1,25	istuma	742	1,21
kogu	298	1,18	olnud	164	1,24	laps	719	1,17
mõte	296	1,17	vana	163	1,23	nagu	705	1,15
väga	292	1,15	uks	159	1,21	isegi	683	1,12
uus	290	1,15	poiss	157	1,19	väga	681	1,11
kaks	287	1,14	arvama	157	1,19	mõte	675	1,10

Ajakirjandustekstide sõnad								
1930			1960			1990		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
olema	4458	38,10	ja	6889	34,27	olema	16228	42,17
ja	3476	29,71	olema	6317	31,43	ja	10881	28,28
see	2328	19,90	see	3698	18,40	see	7618	19,80
tema	1370	11,71	et	1916	9,53	et	4366	11,35
et	1323	11,31	tema	1810	9,00	el	3864	10,04
el	1029	8,79	mina	1548	7,70	mina	3439	8,94
mis	1018	8,70	mis	1506	7,49	tema	3173	8,25
kui	851	7,27	el	1300	6,47	ka	3096	8,05
ka	820	7,01	ka	1181	5,88	Eesti	2968	7,71
oma	681	5,82	nõukogu	1061	5,28	mis	2916	7,58
mina	681	5,82	oma	1012	5,03	kui	2842	7,39
kes	621	5,31	kui	941	4,68	oma	2049	5,32
aga	448	3,83	aasta	880	4,38	saama	1889	4,91
pidama	415	3,55	ning	852	4,24	ning	1739	4,52
siis	408	3,49	töö	839	4,17	aasta	1730	4,50
aasta	402	3,44	kes	804	4,00	aga	1728	4,49
uus	388	3,32	NSV	763	3,80	kes	1667	4,33
tulema	380	3,25	kõik	729	3,63	tulema	1396	3,63
ning	380	3,25	liit	724	3,60	pidama	1387	3,60
üks	370	3,16	suur	674	3,35	või	1260	3,27
saama	364	3,11	Eesti	669	3,33	nii	1235	3,21
teine	363	3,10	saama	637	3,17	võima	1210	3,14
kõik	328	2,80	teine	607	3,02	kõik	1172	3,05
juba	327	2,79	aga	601	2,99	üks	1153	3,00
pool	313	2,68	uus	580	2,89	siis	1143	2,97

Ajakirjandustekstide sõnad								
1930			1960			1990		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
veel	311	2,66	pidama	544	2,71	ise	1126	2,93
suur	311	2,66	võtma	526	2,62	tegema	992	2,58
ise	310	2,65	partei	519	2,58	teine	960	2,49
võima	303	2,59	tulema	515	2,56	aeg	940	2,44
nii	299	2,56	rahvas	509	2,53	veel	856	2,22
aeg	284	2,43	tegema	507	2,52	inimene	815	2,12
kuid	278	2,38	andma	495	2,46	palju	780	2,03
või	268	2,29	vastu	465	2,31	uus	756	1,96
võtma	257	2,20	aeg	452	2,25	võtma	727	1,89
Eesti	255	2,18	üks	430	2,14	riik	727	1,89
vastu	243	2,08	veel	428	2,13	juba	721	1,87
tegema	238	2,03	kommu-nistlik	418	2,08	suur	712	1,85
üle	232	1,98	palju	408	2,03	andma	711	1,85
ainult	231	1,97	juba	404	2,01	jääma	692	1,80
andma	229	1,96	kuid	402	2,00	kuid	687	1,79
kuna	228	1,95	nii	385	1,92	kas	686	1,78
nüüd	225	1,92	eest	382	1,90	välja	683	1,77
nagu	224	1,91	inimene	379	1,89	kus	658	1,71
osa	217	1,85	esimene	378	1,88	vaid	657	1,71
kr	210	1,79	nagu	368	1,84	ütlema	656	1,71
Tallinn	205	1,75	üle	363	1,81	pool	638	1,66
kus	201	1,72	vabariik	361	1,80	nagu	628	1,63
välja	197	1,68	rajoon	357	1,78	minema	625	1,62
vaid	195	1,67	Tallinn	356	1,77	üle	623	1,62
mees	188	1,61	siis	355	1,77	Tallinn	620	1,61
esimene	188	1,61	kesk-komitee	354	1,76	töö	615	1,60
kaks	183	1,56	võima	348	1,73	hakkama	598	1,55
sina	181	1,55	ise	344	1,71	esimene	597	1,55
päev	180	1,54	maa	340	1,69	mitte	594	1,54
palju	180	1,54	noor	337	1,68	kaks	573	1,49
mõni	180	1,54	kolhoos	336	1,67	kord	559	1,45
inimene	179	1,53	pool	331	1,65	vastu	554	1,44
minema	178	1,52	osa	331	1,65	eest	532	1,38
praegu	171	1,46	ütlema	330	1,64	mõni	520	1,35
jääma	164	1,40	töötaja	328	1,63	iga	520	1,35
kord	162	1,38	või	327	1,63	väga	512	1,33
mitte	161	1,38	kogu	326	1,62	nüüd	506	1,31
eest	158	1,35	iga	317	1,58	valitsus	504	1,31
peale	157	1,34	ainult	311	1,55	olnud	502	1,30
ütlema	155	1,32	välja	302	1,51	liit	494	1,28
viimane	152	1,30	riik	291	1,45	päev	493	1,28
valitsus	151	1,29	valitsus	286	1,42	sina	481	1,25
sest	147	1,26	päev	286	1,42	pärast	480	1,25
Itaalia	147	1,26	sotsia-listlik	272	1,35	selline	479	1,24
linn	146	1,25	viimane	269	1,34	praegu	479	1,24

Ajakirjandustekstide sõnad								
1930			1960			1990		
Sõna	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**
töö	144	1,23	kus	269	1,34	ainult	477	1,24
tahtma	143	1,22	küsimus	265	1,32	mees	469	1,22
küsimus	143	1,22	rahvus-vaheline	254	1,26	osa	466	1,21
suurem	141	1,21	kord	253	1,26	arvama	465	1,21
pärast	141	1,21	ülesanne	252	1,25	sest	463	1,20
kogu	137	1,17	esimees	250	1,24	asi	463	1,20
väga	136	1,16	kaks	249	1,24	vabariik	444	1,15
riik	136	1,16	jääma	249	1,24	kõige	440	1,14
ette	136	1,16	Hiina	244	1,21	laps	435	1,13
iga	135	1,15	ameerika	242	1,20	raha	432	1,12
olnud	133	1,14	praegu	238	1,19	miski	430	1,12
leidma	133	1,14	kongress	237	1,18	küll	427	1,11
kas	133	1,14	kõige	236	1,17	tahtma	425	1,10
järele	133	1,14	ülem-nõukogu	234	1,16	rahvas	424	1,10
naine	132	1,13	pärast	230	1,14	nõukogu	423	1,10
kohta	129	1,10	TASS	229	1,14	viimane	413	1,07
seal	125	1,07	sekretär	229	1,14	maa	413	1,07
olev	125	1,07	mõni	227	1,13	enam	412	1,07
ETA	125	1,07	kool	227	1,13	teadma	404	1,05
asi	124	1,06	mitte	225	1,12	kuidas	394	1,02
nägema	123	1,05	minema	224	1,11	kohta	389	1,01
tagasi	122	1,04	linn	222	1,10	küsimus	388	1,01
enam	122	1,04	vaid	218	1,08	Tartu	385	1,00
vana	121	1,03	ettevõte	218	1,08	linn	384	1,00
rahvas	121	1,03	liige	217	1,08	kogu	384	1,00
sama	120	1,03	NLKP	215	1,07	Pärnu	383	1,00
arvama	120	1,03	kohta	215	1,07	sõna	371	0,96
Soome	115	0,98	sm	212	1,05	tagasi	369	0,96
juures	114	0,97	Tartu	203	1,01	sama	367	0,95
Tartu	113	0,97	nüüd	202	1,00	kroon	364	0,95

Süntaktiline märgendamine – arvutiga ja käsitsi*

Kadri Muischnek, Kaili Müürisep, Heili Orav,
Andriela Rääbis, Heli Uibo
Tartu Ülikool

1. Sissejuhatus

Keelekorpused on juba mõnda aega olnud oluliseks töövahendiks nii lingvistikas kui ka arvutuslingvistikas. Ent korpusest on kasu ainult siis, kui on võimalik sellest vajalikku informatsiooni kätte saada. Leksikograafilist infot on tavaliselt võimalik hankida ka töötlemata korpusest (kuigi täpsema teabe saamiseks on mõistlik kasutada lausestatud ja morfoloogiliselt märgendatud korpust). Kuid sageli peab vajaliku teabe ammutamiseks alustama info lisamisest korpusesse. Seega: kui soovitakse, et korpus ei jääks ainult elektrooniliste tekstide arhiiviks, tuleb tekstidele lisada andmed nende ülesehituse kohta, samuti morfoloogilise ja süntaktilise analüüsi tulemused. Sellist interpretatiivse info lisamist suulist või kirjalikku keelt esindavasse keelekorpusesse nimetatakse märgendamiseks.

Käesolev artikkel käsitleb süntaktiliselt märgendatud eesti keele korpuse loomise algusetappi. Korpuse loomise tingis test- ja treeningmaterjali vajadus eesti keele kitsenduste grammatikal põhineva süntaksianalüsaatori (edaspidi ESTKG) jaoks.

Selline loomuliku teksti süntaktiline analüüs, mille puhul iga lauses esinev sõnavorm peab saama vähemalt ühe ja ideaalis ainsa tõlgenduse, toob välja olemasolevate keelekirjelduste ebamäärasused ja mitmetitõlgendatavused, millele ongi käesolevas artiklis tähelepanu pööratud. Eesti keele süntaktiline struktuur on vähem läbi uuritud kui morfoloogiline ja enne kui jõuame süntaksi osas selgi määral formaliseeritud käsitlusele kui morfoloogias, tuleb lahendada veel mitmeid probleeme, nii lingvistilisi kui ka arvutuslingvistilisi.

* Täname professor Mati Ereltit lahkete nõuannete eest.

1.1. Milleks süntaktiliselt märgendatud korpus?

Sajaprotsendilise täpsuse ja korrektsusega ühestatud testkorpuse loomine on korralikult töötava eesti keele automaatse süntaksianalüsaatori loomise seisukohalt hädavajalik. Meie poolt väljatöötatav analüsaator põhineb kitsenduste grammatika formalismil ja kasutab ühestamiseks inimese poolt koostatud reegleid, nn kitsendusi. Algselt võis materjali kitsenduste väljatöötamiseks saada eksisteerivatest grammatikakirjeldustest, põhiliselt "Eesti keele grammatikast" (edaspidi EKG (EKG I 1995, EKG II 1993)), oli olemas ka 16 000-sõnaline käsitsi süntaktiliselt analüüsitud testkorpus, mida kasutati reeglite testimiseks ja programmi töö hindamiseks. Kuid oli selge, et programmi tööd pole enam võimalik oluliselt parandada ilma suurema ja mitmekülgsema testkorpuseta. Grammatikakirjeldused olid ennast uute reeglite allikana amendanud, oli vaja tegelikku keelematerjali, suuremat tekstihulka, millele juba olemasolevaid reegleid rakendades saaks leida need tüüpjuhud, mille süntaksianalüsaator jätab mitteüheseks või analüüsib valesti. Olemasoleva testkorpuse puuduseks oli ka see, et ta koosnes ainult ilukirjandustekstidest. Suurema testkorpuse koostamisel võeti sinna lisaks ilukirjandusele ka ajalehetekste ja juriidilisi tekste.

1999. aastal ühestati käsitsi järgmised tekstid (kokku 24 000 sõna):

- väljavõte 1995–1996. a ajalehtede korpusest – 10 000 sõna;
- ilukirjandustekste Tartu Ülikooli eesti kirjakeele baaskorpusest (1980. aastate korpusest) – 6000 sõna ja 2000 sõna G. Orwelli romaanist "1984";
- juriidilisi tekste 6000 sõna (Isikut tõendavate dokumentide seadus ja Vabariigi presidendi valimise seadus).

Test- ja treeningkorpuse loomist toetati 1999. aastal "Eesti keeletehnoloogia sihtprogrammist" (Eesti Informaatikakeskuse leping nr 915/2404/R2-2/LMTAT0399 "Eesti keele testkorpus" projekti juht Tiit Roosmaa).

1.2. Kitsenduste grammatikal põhinev eesti keele süntaksianalüsaator

Eesti keele kitsenduste grammatika süntaksianalüsaatori ammendava kirjelduse leiab Kaili Müürisepa doktoritööst (Müürisep 2000),

lühema ülevaate annab ka artikkel ajakirjas "Keel ja Kirjandus" (Müürisep 1998). Siinkohal peatume põgusalt selle analüsaatori olulisematel omadustel.

Kitsenduste grammatika töötati välja Helsingi ülikoolis, idee autoriks on prof. Fred Karlsson. Selle formalismi põhjaliku käsitluse leiab tema teosest (Karlsson jt 1995).

Kitsenduste grammatika lähenemine morfoloogilisele ja süntaktilisele analüüsile põhineb kahel traditsioonilisel seisukohal. Esiteks: keel on avatud süsteem, kus ei ole kindlat piiri grammatiliste ja ebagrammatiliste lausete vahel. Seetõttu ei vasta ükski grammatika täielikult tegelikule keelekasutusele. Teiseks: süntaksi aluseks on morfoloogia, eriti morfoloogiliste tunnuste keelespetsiifilised süsteemid. Süntaktilised reeglid so. süntaksianalüüsi reeglid on üldistused, mis kirjeldavad

a) kuidas sõnavormid, mida kirjeldatakse morfoloogiliste tunnuste kompleksidena, avalduvad teatud sõnajärjena;
b) milliseid loomulikke klasse so. süntaktilisi funktsioone saab eristada ja järeldada sellistest sõnajärgedest (Karlsson jt 1995: 37).

Kitsenduste grammatika esmane eesmärk ei ole mitte väljendada või kirjeldada süntaktilisi nähtusi maksimaalselt väheste, unitaarsete, abstraktsete ja komplekssete üldistustena. Pigem võib selle formalismi raames väljendada sama nähtuse erinevaid tahke mitme erineva kitsendusega. Sagedamini kui teistel reeglipõhistel ühestajatel on kitsenduste grammatikal reeglite objektiks üksiksõnad või morfoloogilised tunnused ühekaupa. Seega on kitsenduste grammatika põhiline eesmärk aidata luua mõistliku kirjeldusega ja praktiliselt tõhusaid süntaksianalüsaatoreid, mis baseeruvad ainult pindsüntaksil. Kitsenduste grammatika lähenemine teksti morfosüntaktilisele struktuurile on lähedane traditsioonilisele süntaksile, mille tuumaks on sõnamuutmine, ühildumine ja sõnajärg (Karlsson jt 1995: 38).

Kitsenduste grammatika on oma loomult reduktsionistlik, so analüüsi alguses lisatakse igale sõnale kõik võimalikud analüüsid ja siis hakatakse neid reeglite abil eemaldama. Seetõttu nimetataksegi selles formalismis kasutatavaid reegleid kitsendusteks. Kui "kindlad" reeglid ehk kitsendused pole suutnud teksti ühestada, kasutatakse tõenäosuslikke reegleid, kuid formalismi autorid rõhutavad, et kitsenduste grammatika tuumaks on eelkõige lingvistilised reeglid. "Kindel" reegel võiks olla näiteks selline: kui selle osalause

öeldisverb on umbisikulises tegumoes, kustuta aluse tõlgendused selles osalauses. Ja vastavalt sobiks tõenäosusliku reegli näiteks: kui osalauses on kaks aluse ja öeldistäite kandidaati, siis ühesta neist esimene aluseks ja teine öeldistäiteks.

Süntaktilise analüüsi väljundiks on tekst, milles igale sõnale on lisatud tema süntaktiline funktsioon tekstis. Süntaksipuid ja teisi hierarhilisi struktuure ei genereerita. Formalismi autorid deklareerivad, et kuigi nende eesmärgiks on perfektne analüüs, jätab nende analüsaator pigem ühestamata sellised sõnavormid, millele antud kontekstis ongi omane ebamäärasus või mitmetähenduslikkus. Analüsaator peab suutma anda mingi analüüsi igale sisendile, so. peab olema suuteline analüüsima ka mittetäielikke või vigaseid lauseid. Kitsenduste grammatika formalism on keelest sõltumatu, so. sama formalismi baasil saab luua reegleid erinevate keelte jaoks. Kitsenduste grammatika reegleid ja leksikoni saab kohandada erinevatele tekstitüüpidele, mis on oluline süntaksianalüsaatoril baseeruvate praktiliste rakenduste väljatöötamisel.

Kitsenduste grammatika formalismil põhinevaid analüsaatoreid on loodud lisaks inglise ja eesti keeltele näiteks ka baski, norra, portugali, rootsi ja türgi keelte jaoks (lähemalt vt Müürisep 2000 ptk 2.10).

Eesti keele analüüsiks kasutatakse kahte iseseisvat kitsenduste grammatika formalismil põhinevat ühestajat – eesti keele kitsenduste grammatika morfoloogiline ühestaja (Puolakainen 1998, 2000) ja süntaksianalüsaator (Müürisep 1998, 2000).

Eesti keele kitsenduste grammatika süntaksianalüsaatoris kasutatakse kaht tüüpi süntaktilisi märgendeid (vt tabel 1). Ühed on fraasi põhja märgendid, nagu subjekt, objekt, predikatiiv, öeldisverb, adverbiaal. Teised on laiendite märgendid, mis näitavad ka laiendi asendit põhisõna suhtes.

Tabel 1. Eesti keele kitsenduste grammatika süntaksianalüsaatori märgendid

@+FMV	finiitne öeldis
@-FMV	infiniitne öeldis
@+FCV	*olema* liitaegades, modaalverbid jt ahelverbides, finiitne vorm
@-FCV	*olema* liitaegades, modaalverbid jt ahelverbides, infiniitne vorm
@NEG	verbi eitus (*ei*)
@SUBJ	Subjekt

@OBJ	Objekt
@PRD	Predikatiiv
@ADVL	Adverbiaal
@<Q	kvantori järellaiend (*kaks meest [@<Q]*)
Q>@	kvantori eeslaiend (*inimesi [@Q>] tulvil*)
@<P	eessõnafraasi kuuluv käändsõna (*enne [@<P] õhtut*
@P>	tagasõnafraasi kuuluv käändsõna (*maja [@P>] taga*)
@AN>	adjektiiv eestäiendina
@<AN	adjektiiv järeltäiendina
@VN>	partitsiip eestäiendina
@<VN	partitsiip järeltäiendina
@NN>	nimisõna eestäiendina
@<NN	nimisõna järeltäiendina
@PN>	kaassõna eestäiendina
@<PN	kaassõna järeltäiendina
@INF_N>	infinitiiv eestäiendina
@<INF_N	infinitiiv järeltäiendina

Eesti keele kitsenduste grammatika süntaks põhineb EKG II osal (EKG II 1993), kuid on ka mõned väikesed erinevused.

EKG lubab öeldistäiteks analüüsida teatud juhtudel ka nimisõnu kaasa- või ilmaütlevas käändes (*tüdruk on patsidega*), ESTKG analüüsib sellised sõnavormid määrusteks.

Ka ei tee süntaksianalüsaator vahet fraasiadverbiaalil ja lauseadverbiaalil. Lause- ja fraasilaiendid saavad mõlemad adverbiaali analüüsi.

EKG II loeb öeldise osaks ka ühend- ja väljendverbi määr- ja nimisõnalised komponendid, aga ESTKG analüüsib öeldise koosseisu kuuluvaid nimi- ja määrsõnu eraldi. Enamasti saab ühendverbi koosseisus olev määrsõna adverbiaali tõlgenduse, väljendverbi koosseisus olev nimisõna saab oma grammatilisele vormile vastava analüüsi. Nende öeldise komponentide täpsem analüüs nõuab semantilise info olemasolu ja mahukat väljend- ja ühendverbide arvutisõnastikku. Tõsi küll, esialgne ühendverbide sõnastik, mis baseerub Paul Saagpaku "Eesti–inglise sõnaraamatul" (Saagpakk 1992) ja "Eesti kirjakeele seletussõnaraamatul" (EKSS 1988–1997), on valmimas. Saagpaku sõnaraamatul baseerub ka intransitiivsete verbide sõnastik, mis on edukalt reeglistikuga liidetud. Näiteks juhul, kui öeldisverb on intransitiivne ja lauses pole ka teisi sihilisi

verbe, kustutatakse selle osalause piires kõik objekti tõlgendused. Peab arvestama, et verbide ebatüüpilised kasutused (nt *magab rasket und*) tekitavad siiski vigu, kuid vigade osakaal on võrreldes objekti märgendi eemaldamise efektiivsusega tühine (antud reegel põhjustas kaks viga 30 000-sõnalises korpuses: *astus sammu edasi; Sammu, mille astus Eesti, ...* kus eemaldati ekslikult objekti tõlgendus).

EKG eristab täis- ja osaalust, täis- ja osasihitist ning täis- ja osaöeldistäidet. Eesti keele kitsenduste grammatika koostamisel ei osutunud süntaktilise funktsiooni sisene eristus vajalikuks, sest selle informatsiooni annab morfoloogiline märgend (kääne).

EKG eristab adjektiivatribuuti, genitiivatribuuti ning adverbiaalatribuuti, eesti keele kitsenduste grammatikas ei ole see võimalik, sest selline eristamine eeldaks semantilise info olemasolu. Apositsiooni märgendatakse nagu tavalist nimisõnalist atribuuti.

Kitsenduste grammatika formalism eeldab lause analüüsi ühe sõna kaupa ja mingit fraasistruktuuri ei moodustata. Kaassõnafraasi põhi (so kaassõna) märgendatakse kogu fraasi funktsiooni näitava märgendiga ja kaassõnafraasi kuuluvad kaassõnast sõltuvad nimisõnad märgendiga @P> või @<P. Samuti märgendatakse kvantorifraasi põhi (so kvantor) kogu fraasi funktsiooni näitava märgendiga ja kvantori laiendid märgendiga @Q> või @<Q. EKG järgi on käändsõnalise kvantori laiend täiend, aga määrsõna laiend sõltuvusmäärus. Eesti keele kitsenduste grammatika märgendab mõlemat tüüpi kvantori laiendeid märgendiga @Q, sest kvantori laiendi kohta kehtivad teistsugused reeglid kui tavaliste täiendite ja määruste kohta (lähemalt vt 2.2.6).

Nooled märgendite juures näitavad, kummal pool asub fraasi põhi. Seega, kuigi kitsenduste grammatika süntaksianalüsaator ei anna väljundina lause fraasistruktuuri, on selliselt analüüsitud tekstis siiski võimalik osaliselt fraasistruktuuri tuvastada.

Kitsenduste grammatika formalismi kriitikud väidavad, et kitsenduste grammatika analüüs jääb siiski morfoloogilise ja süntaktilise analüüsi vahepealseks, kuna fraasistruktuuri ei väljastata.

Kitsenduste grammatika järglane, sõltuvusgrammatika (*Dependency Grammar*, vt Järvinen, Tapanainen 1997) analüsaator, on võimeline leidma täpseid sõnadevahelisi süntaktilisi sõltuvusi, mis näitavad laiendi ja fraasipõhja suhet. Iga viide on varustatud märgendiga, mis näitab laiendi süntaktilist funktsiooni. Graafiliselt kujutatuna on analüüsi väljundiks sõltuvuspuu.

2. Vead ja erimeelsused eestikeelse teksti süntaktilisel analüüsil

2.1. Millest nad tekivad?

Süntaktiliseks analüüsiks (täpsemalt: ühestamiseks) tuleb tekstid kõigepealt morfoloogiliselt analüüsida ja ühestada. Kuna eesmärgiks oli luua sajaprotsendilise korrektsusega analüüsitud test- ja treeningkorpus, siis otsustati ka morfoloogiline ühestamine teha käsitsi. Morfoloogilisel analüüsil kasutati morfoloogiaanalüsaatorit ESTMORF koos oletajaga (Kaalep 1998; Kaalep, Vaino 2000), selle väljund ühestati käsitsi. (Käsitsi morfoloogilise ühestamise probleemidest vt Kaalep jt 2000.) Kuna morfoloogiliselt ühestas teksti ainult üks inimene, lipsas ikkagi sisse mõningaid näpuvigu, mis hiljem süntaktilise ühestamise käigus parandamist nõudsid.

Morfoloogiliselt ühestatud tekstidele lisati kitsenduste grammatika süntaksianalüsaatori abil kõik antud sõnavormi puhul võimalikud süntaktilised tõlgendused. Seejärel lasti tekst läbi süntaktilise ühestaja, kasutades kõige kindlamaid ühestamise kitsendusi so selliseid reegleid, mille kohta eelneva kogemuse põhjal oli teada, et need teevad kõige vähem vigu. Ühestaja väljundit analüüsiti käsitsi, kusjuures tähelepanu ei pööratud mitte ainult analüsaatori poolt pakutud märgendite hulgast õige väljavalimisele, vaid üle vaadati ka need sõnad, millel juba oli ainult üks analüüs. Ühestamisel oli aluseks Kaili Müürisepa koostatud juhend (http://www.cs.ut.ee/~kaili/syntax.html), vajadusel otsiti abi "Eesti keele grammatikast" (EKG II 1993).

Sellist sajaprotsendilise täpsuse ja korrektsusega analüüsitud testkorpust on võimalik süntaktiliselt ühestada ainult käsitsi, kusjuures töö käigus selgus, et ka kaks eesti filoloogi haridusega inimest võivad sama konstruktsiooni analüüsida erinevalt. Selle põhjuseks on nii inimlikud vead, EKG süntaksikirjelduse mõningane üldsõnalisus, aga ka asjaolu, et nii mõnedki laused ongi oma süntaktiliselt struktuurilt ambivalentsed, neid ongi võimalik analüüsida mitmel viisil. Seetõttu otsustati, et testkorpuse loomisel tuleb kõiki tekste analüüsida kahe erineva lingvisti poolt, seejärel tuvastada erinevused tulemuses ja arutada need läbi kogu ühestajate grupiga.

Kui sõnavormi morfoloogilise tõlgenduse üle otsustamisel tuleb aeg-ajalt arvestada ka selle sõnavormi süntaktilist funktsiooni (vt ka

Kaalep jt 2000), siis süntaktilise funktsiooni määrab küllaltki tihti semantika. Automaatse süntaksianalüsaatori koostamisel saab semantikat arvestada ainult leksikonide abil. Kuid mõnikord pole nendestki abi – sama sõna võib samas kontekstis siiski kanda erinevat tähendusvarjundit ja peaks seetõttu saama erineva süntaktilise analüüsi. Eriti raske on anda formaalset definitsiooni infiniitsele öeldisele – nii on supiin kohamääruseks lauses *Ta läks jalutama,* aga öeldise infiniitseks komponendiks lauses *Maja läks põlema.* Käsitsi märgendajad lähtusid sellistel juhtudel oma arusaamisest fraasi tähendusest, arvuti jaoks on koostatud verbide leksikon. Kuid mitmetähenduslike verbide korral jääbki automaatsel analüüsil palju mitmesusi.

Grammatikakirjeldused kirjeldavad tüüpiliselt vaid täiesti õigekeelseid juhtumeid, kõnekeelsemate fraaside kohta märgitakse parimal juhul, et need on ebasoovitavad, mõnede väljendite kohta võib väita, et need on fraasistunud ja lauseliikmeid neis määrata ei saa. Kuid tekstikorpuse analüüsimisel ei saa teatud lauseid või fraase analüüsimata jätta, mingi analüüsi peab saama igasugune seotud tekst. Erinevat tüüpi tekstide süntaktilise analüüsi kogemuse põhjal võib väita, et selles suhtes kõige keerukamateks osutusid ajakirjandustekstid, kus leidus nii kõnekeelseid konstruktsioone kui ka märksa keerukamaid lausekonstruktsioone võrreldes ilukirjandustekstidega.

Selline näeb välja süntaktiliselt ühestatud lause meie test- ja treeningkorpuses. Õige analüüs on märgitud plussmärgiga.

Reisidokument
reisi_dokument+0 //_S_ com sg nom #cap // **CLB-C +@SUBJ @PRD
on
ole+0 //_V_ main indic pres ps3 sg ps af #FinV #Intr // @+FMV
seaduses
seadus+s //_S_ com sg in // +@ADVL @NN>
riigipiiri
riigi_piir+0 //_S_ com sg gen // @NN>
ületamiseks
ületa=mine+ks //_S_ com sg tr #mine // +@ADVL @<NN
ettenähtud
ette_nähtud+0 //_A_ pos // @AN>
Eesti
Eesti+0 //_S_ prop sg gen #cap // @NN>
dokument
dokument+0 //_S_ com sg nom // @SUBJ +@PRD @NN> @<NN
või
või+0 //_J_ crd // @J

Välisministeeriumi
välis_ministeerium+0 //_S_ com sg gen #cap // @P>
poolt
poolt+0 //_K_ post #gen // +@ADVL @PN> @<PN
tunnustatud
tunnustatud+0 //_A_ pos // @AN>
välisriigi
välis_riik+0 //_S_ com sg gen // @NN>
reisidokument
reisi_dokument+0 //_S_ com sg nom // @SUBJ +@PRD @NN> @<NN
$(
$(//_Z_ Opr //
edaspidi
edas_pidi+0 //_D_ // +@ADVL @AD>
välisriigi
välis_riik+0 //_S_ com sg gen // @NN>
reisidokument
reisi_dokument+0 //_S_ com sg nom // +@PRD @<NN
$)
$) //_Z_ Cpr //
$.
$. //_Z_ Fst //

2.2. Tüüpilisemad erimeelsused

Allpool käsitleme süntaktilisi funktsioone, mille analüüsimisel olid sama teksti ühestanud lingvistidel enam erimeelsusi. Tabelis 2 on ära toodud test- ja treeningkorpuse analüüsimisel tekkinud vigade liigitus, ära on toodud ka tehnilised vead ja morfoloogilise ühestamise vead. Tehnilisi vigu võis olla juba sisendtekstis, neid põhjustasid morfoloogiline analüsaator, teisendusprogrammid ja süntaksianalüsaator. Nagu öeldud, ühestati tekstid eelnevalt käsitsi ka morfoloogiliselt. Nii on morfoloogilise ühestamise vead nn inimlikud vead.

Tabel 2. Põhilised vead ja erimeelsused süntaktilisel ühestamisel

da-infinitiiv	36
Noomenifraas	85
Määrus	208
Passiiv	32
Kvantorifraas	38
Muud	187
Kokku	**586**
Morfoloogilise ühestamise vead	116
Tehnilised vead	186
Kõik kokku	**888**

2.2.1. Substantiivifraas

Substantiivifraasiks nimetatakse fraasi, mille põhjaks on substantiiv, ka pro- või kvaasisubstantiiv. Substantiivi laiendid on atribuut ning apositsioon. Nagu öeldud, ESTKG atribuuti ja apositsiooni ei erista, lisandit märgendatakse nagu tavalist nimisõnalist täiendit. Samuti ei eristata adjektiivatribuuti, genitiivatribuuti ning adverbiaalatribuuti, kuna see eeldaks semantilise info olemasolu.

Fraasi põhja määramisel tekkis probleeme, kui fraas koosnes kahest asesõnast. Lahkarvamusi tekitasid näiteks fraasid *see kõik*, *kõik muu*, *keegi teine*, *midagi säärast*, *midagi muud*.

Asesõnafraasides, mille moodustavad demonstratiiv- ja/või indefiniitpronoomenid, on tarindis esimesel kohal paiknev pronoomen täiendiks, teine substantiivselt tarvitatuna põhjaks (EKG II 1993: 118). Lauses *Ta tundis oma naist lähemalt kui kedagi teist* on *kedagi* täiend ning *teist* põhi. Pronoomen *miski* on aga indefiniitpronoomenite hulgas erandlik: ta on põhijuhul substantiivne ning ei talitle adjektiivatribuudina, vaid põhjana. Lauses ... *oli raske ette kujutada, et tegelikkuses midagi säärast ette tuleb* on *midagi* alus ning *säärast* täiend, samuti nagu lauses *Segadus maailma suuruselt teises tuumariigis on midagi muud kui pelgalt akadeemilist huvi pakkuv purelemine Bütsantsi õukonnas* on *midagi* öeldistäide ning *muud* järeltäiend.

Pronoomen *kõik* esineb nii ees- kui järelasendis. Lauses *Eesti sai selle kõik peaaegu kahe aastaga* on *selle* sihitis ning *kõik* järeltäiend. Lauses *Kõik muu oli Jürile varasemastki tuttav* on *kõik* aga eestäiend ning *muu* alus.

Determinatiivpronoomen *ise* esineb põhiliselt järeltäiendi positsioonis, nt *Marinale endale läks keelepeks vähe korda*. Probleeme tekkis eksplitsiitse subjekti puudumise korral, nt *otsustasin ka ise võtta suuna*. Ka niisugusel juhul on *ise* täiend, kuigi põhi puudub.

Raskusi põhjustas ka mõne asesõnast ja arvsõnast koosneva fraasi põhja määramine, nt *neid kahte vaadates*; *kedagi kolmandat otsima hakata*. Niisugustes fraasides on asesõna täiend ning arvsõna põhi.

Aega märkivad nimisõnafraasid on EKG järgi tervikuna ajamäärused. ESTKG analüüsib lauset aga sõnade, mitte fraaside kaupa. Nt fraasi *15. veebruaril 1999* peasõnaks määrati kokku-

leppeliselt *veebruaril*, *15.* ja *1999* on täiendid; fraasi *1999. aasta veebruaris* põhi on *veebruaris* ning täiendid *1999.* ning *aasta.*

Lisand on raskemini tuvastatav kui täiend, kuna nii lisand kui ka tema põhi on substantiivifraasid ja lisand võib paikneda nii põhja ees kui järel.

Üksikuid vigu tehti fraaside puhul, mille põhjaks on isikunimi ning laiendiks põhja täpsustav eeslisand. Mitmeosalistes isikunimedes loetakse esimene tarindiliige lisandiks. Nt fraasis *prokurör Urmas Tammiksaar* on *prokurör* ja *Urmas* lisandid ning *Tammiksaar* põhi.

Firmanimed, mida ajakirjandustekstides esineb palju, tekitasid märksa enam probleeme, nt lauses *Seli on ASi Rondam omanik, ASi Estiko ja ASi Wermo suuraktsionär, ASi Salvest ja Haapsalu hotelli aktsionär.* Niisugustel juhtudel lähtuti sellest, et kui apositsioonitarind koosneb muudetavast ja muutumatust elemendist, siis on muudetav element põhi ning muutumatu element lisand. Fraaside põhjad (öeldistäited) on *omanik*, *suuraktsionär* ning *aktsionär.* Lühendid *ASi* on omakorda neid laiendavad genitiivatribuudid ning apositsioonitarindite põhjad. Tarindite muutumatud elemendid *Rondam*, *Estiko*, *Wermo* ja *Salvest* on järellisandid.

Veel keerulisem oli võõrkeelsete nimede süntaktiline analüüs. EKG järgi on mitmesõnalised jutumärgistatavad nimed järellisandid, ESTKG nõudis aga iga sõna eraldi märgendamist. Lauses *Rootsi äriregistris on küll registreeritud TEWI Group ja selle õigusjärglane Ewizonen i Värnamo Ab* on objekti positsioonis kaks substantiivifraasi: *TEWI Group* ning *selle õigusjärglane Ewizonen i Värnamo Ab.* Esimeses fraasis on *TEWI* lisand ning *Group* põhi. Ka teise fraasi puhul otsustati märkida põhjaks viimane tarindiliige *Ab*, *selle* on täiend ning ülejäänud sõnad lisandid. Kaaluti ka võimalust märgendada *i* sidendiks, kuid sellest siiski loobuti. Tulevikus oleks ilmselt mõtet esitada pikad võõrkeelsed nimed ühe tõlgendusena, kuid selleks tuleb tekstide eeltöötlust täiendada.

Mõnel juhul polnud võimalik üheselt määrata, mille juurde laiend kuulub. Nt lauset *keegi meist tegutses Eesti Vabariigi ja tema kodanike huvide vastu* on võimalik analüüsida kahel viisil: *tegutses Eesti vabariigi vastu*, st *vabariigi* on kaassõna laiend, või *tegutses Eesti vabariigi huvide vastu*, st *vabariigi* on substantiivi laiend. Sellistele lausetele jäeti mitu märgendit.

2.2.2. Määruse analüüsiga seotud probleeme

Määrus ehk adverbiaal on verbi laiend, mis pole alus, sihitis ega öeldistäide. Määrusel puudub selgepiiriline funktsioon lauses, tema vorm oleneb sellest konkreetsest rollist, mida tema referent täidab lausega tähistatavas situatsioonis. Sellest tuleneb määrusevormide rohkus. Määruse vormi ja tema süntaktilise käitumise määrab olulises osas tema tähendus. Nii eristatakse aja-, koha, viisi-, seisundivm määrust. ESTKG siiski ei eristata erinevaid määruse liike.

Määrus ja täiend. Sageli ei suuda ESTKG eristada määrust ja adverbiaalset täiendit. Kahe nimisõna kõrvuti asetsemise korral on väga raske automaatselt kindlaks määrata, kumb on kumma täiend või on nad mõlemad määrused. Mõnikord oli seda raske teha ka käsitsi ühestamisel, sest ei suudetud üheselt otsustada, kas antud fraas kuulub nimisõna või verbi juurde. Suurimaid probleeme põhjustas määruse eristamine nimisõnalistest ees- ja järeltäienditest. Eriti puudutas see kohamäärustena funktsioneerida võivaid sõnu. Lahendusena tuli sageli kasutada sõnavormi mitmeseks jätmist so jäeti alles nii määruse kui täiendi tõlgendused. Mõningaid iseloomulikumaid näiteid:

> Erakorraline valimine Riigikogus toimub...
> ...oli olnud nende ametiposti suur, sageli piiramatu võim kodumaal ning oluline roll regionaal- ja maailmapoliitikas.
> Vene juhtivad poliitikud pole sellises vormis arvamusi esitanud.

Ka kaassõnafraasi adverbiaali või atribuudi tõlgendus tekitas probleeme. Kaassõnafraasi võiks pidada sõltuvusmääruseks, mis on substantsiaalsete määruste jääkklass ja mille tähendusliigid on väga ähmaste piirjoontega.

Lausetes *Jüril tekkis vimm mägede vastu...* ja *Kaljo Mandrel tekkis huvi rahvameditsiini vastu juba lapsepõlves* jäid kaassõnafraasile (täpsemini küll kaassõnale) alles nii adverbiaali kui atribuudi tõlgendused. Kuid teisalt võib ka tunduda, et kaassõna *vastu* ei ole seotud mitte niivõrd verbi kui just subjektnoomeniga (*vimm mille vastu*).

Määrus ja öeldistäide. Raskusi tekitas kvantorfraasi ühestamine määruse ja öeldistäite tõlgenduste vahel.

Kas järgmistes lausetes on kvantorifraaside puhul primaarne määra, kvantiteedi tähendus ja kvantorid peaksid saama adverbiaali tõlgenduse või võib neid analüüsida predikatiividena?

Viivise suurus on 25 miljonit krooni...
Aktsia käive oli eile börsil 3,8 miljonit...
Meil on vaja täita linnavara müügi ülesanne, mis on 37 miljonit krooni.

See on koht, kus traditsiooniline grammatika ei anna ühest lahendit. See, milline analüüsivariant valida, sõltub suuresti sellest, mida tahame süntaktiliselt analüüsitud tekstiga edasi teha. Loogiline järg süntaktilisele ühestamisele on semantiline ühestamine. Kui analüüsida need kvantorifraasid öeldistäideteks, siis oleks lausetel tähendus

Viivise suurus on mis?/milline?
Käive oli eile börsil mis?/milline?

Kui analüüsime kvantorifraasid adverbiaalideks, siis oleksid lausetel tähendus

Viivise suurus oli kui suur?
Käive oli eile börsil kui suur?

Üheks selliste situatsioonide võimalikuks lahenduseks oleks jätta sellised semantikast sõltuvad mitteühesused lahendamata ja pöörduda nende juurde tagasi pärast teksti esialgset semantilist ühestamist.

2.2.3. *da*-infinitiiviga seotud probleeme

da-infinitiivitarind talitleb lauses põhiliselt seotud laiendina, täites subjekti, objekti ja adverbiaali või vastavate atribuudiliikide, ka predikatiivi süntaktilist funktsiooni. Üksikjuhtudel on da-infinitiiv ka vaba laiend (EKG II 1993: 237). Finiitverbile iseloomulikus iseseisva predikaadi rollis kasutatakse *da*-infinitiivi mõningat tüüpi kõrvallauseis ja kaudse kõneviisi asemel kaudse teatelaadi väljendamiseks (EKG II 1993: 244). Lisaks võib *da*-infinitiiv kuuluda ka perifrastilise verbi või verbivormi koosseisu (EKG II 1993: 246–248).

Süntaksianalüsaator lisab *da*-infinitiivile kõige rohkem märgendeid – 7 (ahelverbi komponent, subjekt, objekt, adverbiaal, predikatiiv ning ees- ja järeltäiend).

***da*-infinitiiv öeldise infiniitse komponendina.** Kokkuvõtteid tehes selgus, et ahelverbi komponendi ja subjekti tõlgendused olid da-infinitiivsete konstruktsioonide analüüsil kõige keerukamad. Sageli eelistasid analüüsijad loobuda ahelverbi komponendi tõlgendusest ja valisid selle asemel subjekti, predikatiivi, objekti või määruse tõlgendused.

Kui *da*-infinitiiv moodustab ahelverbi koos modaalverbidega *võima, tohtima ja saama,* või ka deskriptiivverbiga (*paukus köhida, vehkis käia),* on tema analüüsimine suhteliselt hõlbus, sellega saab hakkama automaatne süntaksianalüsaator ja ei tekkinud sellega suuri probleeme ka käsitsi ühestamisel. Siiski nt lauses *Winston mõtles jälle, et tüdruk ei pruugi olla Mõttepolitsei agent* eelistas üks ühestaja subjekti ja teine ahelverbi komponendi tõlgendust.

Küll aga oli probleeme *olema*-verbiga seostuva da-infinitiivi analüüsimisel. Selle kohta vt ka 2.2.5. Erinevad lingvistid olid erinevatel arvamustel nt lause *...aga seda polnud veel võtta...* analüüsil, kaheldi *da*-infinitiivi subjekti ja ahelverbi osa tõlgenduste vahel ja lõpuks jäädi jäädi viimase juurde.

***da*-infinitiiv subjektina.** *da*-infinitiiv esineb subjektina peamiselt kogejalauseis ja laiendab

- emotsionaalset või füsioloogilist seisundit väljendavaid predikaate;
- sobivust väljendavaid predikaate;
- vajadust või kohustust väljendavaid predikaate;
- suutelisusele viitavaid verbe;
- substantiive, mis võivad (mujal) esineda *da*-infinitiivse atribuudi põhisõnana, eriti soovi, kavatsust, eesmärke väljendavaid substantiive (EKG II 1993: 237–239).

Peamised ühestamise vead tekkisid *da*-infiniitse sõnavormi subjekti ja ahelverbi komponendi tõlgenduste vahel valimisel.

Lauses ... *öeldi, et midagi pole teha ja tuleb arvestada halvimaga* kaheldi aluse ja öeldise osa tõlgenduste vahel ja jäädi aluse tõlgenduse juurde. EKG II leheküljel 247 on seda tüüpi lausete kohta küll öeldud: "Predikaadi koosseisu kuuluvaiks võib lugeda ka muude verbide *da*-infinitiive; viimaste piir subjekti ja adverbiaaliga ei ole siiski eriti terav" ja näitelausete hulgas on ka lause *Tal polnud midagi teha.* Nii et sellist tüüpi lauseis oleks põhimõtteliselt võimalik ka *da*-infinitiivile mitme süntaktilise märgendi jätmise lubamine, kuid just selliste *pole teha* – tüüpi konstruktsioonidega

lauseid esineb eestikeelses tekstis suhteliselt sageli ja kui neid on vähegi võimalik (lubatav) formaalsete kriteeriumite alusel ühestada, tuleks seda ka teha (vt ka 2.2.4.).

Mõned vead tekkisid ka subjekti ja atribuudi tõlgenduste vahel valimisel. Nii nt on lauses ... *on maailma juhtivatel majandusriikidel plaanis investeerida Kesk- ja Ida-Euroopa riikidesse...* *da*-infinitiivi ekslikult tõlgendatud atribuudina (*plaan investeerida*).

***da*-infinitiiv öeldistäitena.** See *da*-infinitiivi süntaktiline funktsioon ei tekitanud käsitsi ühestamisel nii palju probleeme kui subjekti ja ahelverbi osa õige analüüs. Aga siiski kaheldi nt lauses *Aga see oli teada: kui lähed teisi õpetama...* öeldise osa ja öeldistäite (*teada* oleks siis samas tähenduses nagu fraasis *teada asi*) tõlgenduste vahel.

***da*-infinitiiv objektina.** Objekti tõlgendus ei tekitanud eriti palju probleeme. EKG annab loetelu sagedasematest verbidest, mida võib laiendada infiniitobjekt (EKG II 1993: 240). Selle põhjal on automaatsesse süntaksianalüsaatorisse integreeritud leksikon, mille abil ESTKG määrab *da*-infinitiivi objektiks, kui selline verb leidub ja kustutab objekti märgendi, kui verbi ei leidu. Ühestajad eksisid vaid paaril korral, tõlgendades objekti funktsioonis olevat *da*-infinitiivi kas ahelverbi osa või atribuudina.

***da*-infinitiiv määrusena.** *da*-infinitiiv võib esineda lauses otstarbemäärusena, olles sageli asendatav translatiivse teonimega (EKG II 1993: 241).

Käsitsi ühestamisel tekkis vigu *da*-infinitiivi adverbiaali tõlgenduse eristamisel peamiselt predikatiivi, atribuudi ja subjekti tõlgendustest. Lauses ...*viimane aeg on Hansapanga aktsiasse sisse minna* kaheldi predikatiivi (*aeg on minna*), adverbiaali (*aeg on minemiseks*) ja atribuudi (*aeg minna*) tõlgenduste vahel. Kõne alla tuleks ka subjekti tõlgendus (*minna on aeg*). Lõpuks sai see sõnavorm küll üheselt adverbiaali tõlgenduse, kuid me ei püüagi väita, et see ainuõige lahendus oleks. Sarnane probleem kerkis lause ...*kas on mõtet üldse müüa*? puhul.

Lause *Tulla Eesti jaanuarist Kaukaasia eelmäestikku oli sama hea kui paar pimedat talvekuud vahele jätta.* analüüsimisel oli üks ühestaja eelistanud predikatiivi, teine adverbiaali tõlgendust. Siin on siiski selgelt tegemist adverbiaaliga (*sama hea kui jätta*).

***da*-infinitiiv atribuudi funktsioonis.** Tänapäeva eesti keele grammatika (EKG II 1993: 242–243) ütleb *da*-infiniitse atribuudi kohta, et *da*-infinitiiv laiendab substantiive, mis väljendavad:

- soovi, kavatsust või üritust;
- käsku, keeldu, luba;
- tegevussubjekti sisemisi või väliseid ressursse (*julgus, võime, harjumus* jne).

Problemaatilise täiendina esineva da-infinitiivi näiteks võiks tuua lause *See oli teada asi*, mille analüüsimisel kahtlesid mõlemad ühestajad, kas *da*-infinitiivne sõnavorm *teada* võib olla atribuudiks.

Lauses *Ta tõmbab kliendid üle, sest on parem võrk välja pakkuda* võib *da*-infinitiiv olla predikatiiv (*võrk on pakkuda*) või subjektfraasi osa (*võrk pakkumiseks*). Sama probleem oli ka lause *Krunti polnud mõtet loovutada* analüüsimisel. Lause puhul on võimalikud tõlgendused: *mõtet* subjekt ja *loovutada* atribuut või *loovutada* adverbiaal.

2.2.4. *ma*-infinitiiviga konstruktsioonide analüüs

ma-infinitiiviga konstruktsioonide analüüsil tekitas raskusi adverbiaali ja ahelverbi komponendi tõlgenduste eristamine. Nt lausetes

> Marina pani grusiinlastest ülemused oma pilli järgi tantsima.
> ... ta hindas oma elukogemusi ja õpetas meestki nende järgi elama.

oli suuri raskusi *ma*-infinitiivi süntaktilise funktsiooni üle otsustamisel. Ja tõepoolest polegi nii lihtne järgneva kahe definitsiooni põhjal otsustada, millal tuleks kausatiivverbi laiendav *ma*-infinitiiv analüüsida adverbiaaliks ja millal ahelverbi osaks. Nimelt kirjutatakse "Eesti keele grammatikas" et finaalse kohaadverbiaalina talitlev *ma*-tarind laiendab ka järgmisi verbe ja noomeneid, mille tähenduses viide kohamuutusele puudub – kausatiivverbe: *käsutama, hõikama, hüüdma, kutsuma*. Sõltuvusmäärusena laiendab *ma*-infinitiiv eespool toodud 5. rühma kausatiivverbe, neile lisaks ka mittelokatiivse laiendiga seostuvaid kausatiivverbe: *sundima, kohustama, provotseerima* jne (EKG II 1993: 253–254).

Ahelverbi moodustab *ma*-tegevusnimi ka nende verbidega, mis väljendavad protsessi või seisundi kauseerimist (tihti ka kauseerimise laadi): *jätma, panema, ajama* (EKG II 1993: 258). Kui inimene ma-infinitiivis sõnavormi ühestamisega ka toime tuleb, siis kahjuks

automaatne süntaksianalüsaator saab seda teha ainult leksikoni abil ja kui verbi leksikonis pole, siis jääb *ma*-infinitiiv mitmeseks.

Alati on muidugi keeruline väljajätteliste lausete analüüs. Nt lauses *Nüüd sööma, kamandas...* otsustati jätta ma-infinitiivile nii ahelverbi osa kui adverbiaali tõlgendused.

2.2.5. Passiiviga lausete analüüs

Passiivilausetes ei esine tegevussubjekt grammatilise subjektina. Grammatiline subjekt on kas lausest välja jäänud (subjektita passiiv) või on grammatilise subjekti positsioonis hoopis tegevusobjekt (subjektiline passiiv) ning tegevussubjekt esineb kas agentadverbiaalina või puudub lauses üldse (EKG II 1993: 30).

Passiivilausete öeldised võivad olla järgmistes vormides:

- verb umbisikulises tegumoes;
- verb isikulise tegumoe 3. pöördes (mitmus on siiski haruldane);
- verbi *saama* ainsuse 3. pööre + *tud*-partitsiip;
- verbi *olema* ainsuse 3. pööre + *da*-infinitiiv;
- verbi *saama* ainsuse 3. pööre + *da*-infinitiiv.

Kõik need öeldise vormid tekitasid käsitsi analüüsil segadust: kas nimetavas või osastavas käändes sõna on subjekt või objekt, kas *tud*-partitsiip on öeldise osa, predikatiiv või adverbiaal, kas *olema*-verb on iseseisev öeldis või vormistab ainult liitaega.

Verb umbisikulises tegumoes. Umbisikulise tegumoe olevik ja lihtminevik tavaliselt probleeme ei tekita. Leitud vead võib pigem näpuvigadeks lugeda. Subjekti märgendit osatakse ka automaatsel analüüsil vigadeta eemaldada. Nt *Investeering firmasse Starman tehti Eesti Sideministeeriumi informeerimata.* See-eest on umbisikulise tegumoe täis- ning enneminevikku kerge ajada segi seisundipassiiviga, mida vormistab subjekt nominatiivis, verb *olema* ning *tud*-partitsiip. Nt *Pensionär oli liigutatud.* Mõlemal juhul ühildub *olema*-verb nimetavas käändes sõnaga: *Need olid lõplikult sassi aetud.* EKG eristab seisundipasiivi tegelikust passiivist tähenduse põhjal: "kui passiivse lause puhul tähistab tegevussubjekt suhteliselt aktiivset elusat osalist, harilikult agenti või kogejat, siis seisundipasiivi korral tähistab tegevussubjekt seisundi vahetut elutut põhjustajat, mida vormistab seestütlev kääne, kaassõnafraas N_g+*üle* vms." (EKG II 1993: 30). Nt:

... kuid eelkõige olid nende püsimisest huvitatud võimurite kitsamasse või laiemasse lähikonda kuuluvad isikud.

Kuid mõnikord on raske otsustada, kes või mis on tegevussubjekt ning kas *tud*-partitsiip väljendab tegevust või seisundit. Korpuse analüüsil eksiti konstruktsiooniga *olema* + *seotud* neljal korral. Nt *Sinijärve sõnul on Venemaa eksperdi avaldus seotud Venemaal peagi toimuvate valimistega.*

Mõnes lauses on tegevussubjekt kindlasti elusolend, kuid *tud*-partitsiip esineb predikatiivina: *mis oli poest ostetud ja paras hapupiimaks nimetada.*

Kõige kindlamaks testiks *tud*-partitsiibi analüüsil osutus lause eitavaks muutmine. Kui nimetav kääne säilib, on tegemist aluse ja öeldistäitega, vastasel korral sihitise ja öeldisega. Kui võimalikud on mõlemad variandid, jäetaksegi lausesse mitu analüüsi: *kui Prantsuse tuumaarsenali tulevane ohutus ja töökindlus on tagatud.* Kui aga potentsiaalseks aluseks on *da*-infinitiiv, sobib testiks *olema*-verbi muutmine liitvormiks (*on olnud*). Nt *Kui välisleping ei sätesta teisiti, ei ole Eestist lahkumisel lubatud kaitseväe tunnistust kaasa võtta.* Võimalik on öelda: *võtta on olnud lubatud.*

Verb isikulise tegumoe 3. isikus. Passiiv, milles öeldiseks on 3. isikus verb, probleeme ei tekitanud. Sageli on siis sihitis omastavas käändes ning seetõttu on teda alusega võimatu segi ajada. Seda liiki passiiv põhjustas ainult ühe vea, mida võib ka pigem näpuveaks klassifitseerida: *Seda võis vaevalt vanaaegses mõttes koduks nimetada.*

Verb *saama* ainuse 3. pöördes + *tud*-partitsiip. Verb saama ainsuse 3. pöördes koos *tud*-partitsiibiga väljendab umbisikut so. viitab tegevuse olemasolevale, kuid täpsemalt konkretiseerimata elusale sooritajale. Nt *Juba krunt sai võetud mitte iseäralise tarviduse pärast.* Seda tüüpi lausetes on nimetavas käändes nimisõna sihtis ja *tud*-partitsiip öeldise osa. Et abisõna *saama* võib ka täissihitisega ühilduda, analüüsiti paar korda sihitis ekslikult aluseks. Nt *Isegi põhiseadus ja valimisseadus said kirjutatud nende kahe mehe vägikaikaveost lähtuvalt.*

Segadust tekitab ka võimalus analüüsida *tud*-partitsiip seisundimääruseks lauses *Tööd said tehtud* (EKG II 1993: 265). Nii analüüsiti lauses *Pärast seda, kui naine suri, maha maetud sai ja esimene hämmeldus uuest olukorrast vaibuma hakkas ... tud*-partitsiip *maha*

maetud kui määrus (*naine* on ju lause esimeses pooles alus), see aga ei osutunud korrektseks. Testiks sobib samuti lause eitavasse kõnesse panek. Kui nimetav kääne jääb alles, on tegemist alusega ning *tud*-partitsiipi tuleb käsitleda kui seisundimäärust. Nt *Tööd ei saanud tehtud. Kana ei saanud kitkutud. Pärast seda, kui naist ei saanud maha maetud* .... Samuti sobib testiks *saama*-verbi vormi muutmine täis- või enneminevikuks. Kui see on võimalik, on tegu aluse ja seisundimäärusega. Nt *Tööd on saanud tehtud* aga *Naine on saanud maha maetud.*

Verb *olema* ainsuse 3. pööres + *da*-infinitiiv. *olema*-verbi 3. pööre ja *da*-infinitiiv väljendab üldisikut: *Seda on arvata.* Seda tüüpi lausetes on raske just *da*-infinitiivi määramine: kas tegemist on öeldise osa, aluse, määruse või mõne nimisõna täiendiga. Nt *Mis alust oli tal arvata...; Köögis polnud midagi süüa. Krunti polnud mõtet loovutada.*

Verbiga *olema* koos esinev *da*-infinitiiv otsustati analüüsida öeldise osaks, kui *da*-infinitiiv oli üks verbidest *teadma, nägema, märkama, kuulma, tundma, kartma, arvama, lootma, ootama* jne (EKG II 1993: 246). Kuna muude verbide *da*-infinitiive võib samuti lugeda predikaadi koossseisu kuuluvaks ning ka EKG tunnistab, et "viimaste piir subjekti ja adverbiaaliga ei ole eriti terav" (EKG II 1993: 247), siis otsustati muude verbide *da*-infinitiivide korral iga juhtum eraldi. Paljudel juhtudel jäeti mitu märgendamise võimalust.

Verb *saama* + *da*-infinitiiv. Ka need laused, kus *da*-infinitiiv kuulub verbi *saama* juurde, on passiivilaused. Tegevusobjektist grammatiline subjekt ning perifrastiline passiivivorm moodustavad subjektilise passiivi, kus tegevussubjekt võib kas lausest puududa või esineda agentadverbiaalina kujul N_g+*käest* või N_{abl}.

Nt *Oma pahameeleavalduste viisaka leebuse eest sai Uus-Meremaalt ja Austraalialt noomida ka Suurbritannia.* Vigu tekkis siin kahte liiki: aluse asemel analüüsiti sihitis (*Suurbritannia*) või leiti, et *da*-infinitiiv sarnaneb sihtisele (*Ma sain siis teada*), kuid need olid rohkem hajameelsusest tingitud vead.

Kokkuvõtteks võib öelda, et passiivilausetes põhjustas enamiku vigu *tud*-partitsiibi süntaktilise funktsiooni eristamine: kas tegu on öeldise liitajaga või öeldistäitega. Samas õnnestus töö käigus seda tüüpi lausete analüüsireeglid selgeks vaielda ning tulevik peaks selles osas helgem olema. Natuke haruldasema konstruktsiooni

olema + *da*-infinitiiv kohta ei õnnestunud aga kindlaid teste koostada ning ka edaspidi tuleb selle passiivivormi analüüsil lähtuda paljuski eelnevatest näidetest.

2.2.6. Kvantorifraasi probleeme

Kvantorifraas on fraas, mille põhjaks on kvantor ja laiendiks substantiiv või substantiivifraas. Seejuures märgib fraasi laiend mõõdetavat objekti, objektide hulka või ainet, põhi aga selle hulka, mõõtu või määra (EKG II 1993: 140).

Kvantorifraasi põhjaks võib olla

- põhiarvsõna(ühend) või pronumeraal (*kümme sõrme, kolmsada kuuskümmend viis päeva, mitu raamatut*);
- kvantiteedisubstantiiv, sealhulgas mõõtühikud (*meeter riiet, kilogramm juustu, liiter vett, kraad külma*) ja hulka (*kimp lilli, paar sokke*) või osa (*osa peatükke, enamik kohalviibijaist*) märkivad nimisõnad;
- kvantiteediadverb (*palju maiustusi, natuke nalja, rohkem teed*).

Traditsioonilise eesti keele grammatika järgi on noomenist kvantori laiend atribuut, adverbi laiend sõltuvusmäärus. Kitsenduste grammatikas (ESTKG) on nende laiendite jaoks sisse viidud omaette märgend (vt tabel 1). Märgendite @<Q ja @Q> kasutuselevõtt vähendas oluliselt osastavas käändes nimisõnade mitmesust, kuna kvantori laiendit on suhteliselt kerge määrata, kui on antud kvantorite loetelu. Loomulikult ei ole võimalik anda lõplikku nimisõnaliste kvantorite loendit, seetõttu tuleb arvestada, et ESTKG eksib haruldaste kvantorite korral.

Kvantori laiend võib jääda implitsiitseks, nii et eksplitsiitselt esindab fraasi üksnes kvantor. Kvantoriaalne põhi aga peab kindlasti olema eksplitsiitselt väljendatud. Kvantori ärajätmisel tekib kas ebagrammatiline või muutunud tähendusega lause (EKG II 1993: 140).

Süntaksianalüsaatoril on põhimõttelisi probleeme kvantorifraasi laiendite identifitseerimisega. Kui arv on kirjutatud numbritega, siis osastavas käändes nimisõna kuulutatakse kvantori laiendiks. Nt ... *kes teenib 28 miljonit krooni*. Teiste käänete puhul on automaatselt väga raske otsustada, kas järgnev sõna laiendab kvantorit või on ta mõne teise sõna laiendiks. Kui vaadelda meie test- ja treeningkorpuses esinevaid lauseid: ... *kerkis 300 meetri kõrgusele, ... peaks*

kaaluma 51 protsendi müüki, ... viia kaabel 50 000 Tallinna kodusse, Eestis on 62 000 kaabel TV abonenti, siis tundub, et ainsaks lahenduseks on sõnade loetelude tegemine ja heuristilised reeglid. Praegu jäävad paljud arvu järel olevad sõnad mitmeseks. Peamine põhjus, miks kvantorifraasi laiend saab mitmese analüüsi, on arvsõnalise kvantori kirjutamine numbrina, mistõttu ei ole võimalik selle käänet kindlaks määrata (Müürisep 2000: ptk 3.8.5).

Kvantorifraas esineb lauses subjekti, objekti, predikatiivi, adverbiaali või atribuudina. Kvantorifraas võib kuuluda ka omadus-, määrsõna- või kaassõnafraasi koosseisu.

Kvantorifraaside süntaktilise ühestamise probleemid. Analüüsitud tekstilõikudes oli kokku 38 kvantorifraasiga seotud viga. Tunduvalt vähem oli seda tüüpi vigu ilukirjanduses (igas 2000-sõnalises katkendis üks). Seevastu ajalehtedes ja juriidilistes tekstides oli keskmiselt kuus kvantorifraasi viga 2000-sõnalise katkendi kohta. Järgnevalt vaatleme enamlevinud vigade tüüpe.

Osa vigu tuli lihtsalt teadmatusest – põhiarvsõna ja selle laiend olid olemas, aga ühestajal polnud kvantorifraasi määramine veel selge. Isegi see pealtnäha lihtne juhtum, kus fraasi põhjaks on põhiarvsõna ja laiendiks loendatav objekt, võib osutuda keeruliseks, kuna arvsõnafraas võib esineda peale iseseisva lauseliikme ka atribuudina nimisõnafraasis (*15 kilomeetri sõit, 1–2 miili raadiuses*), adverbiaalina omadussõna- ja või määrsõnafraasis (*poolteist meetrit pikk, neljakümne kraadi kangune, viis krooni odavamalt, kümmekond aastat hiljem*) või kaassõnafraasi liikmena (*kümne päeva jooksul, kolmest jõest läbi, üle saja kilo*) (EKG II 1993: 143).

Kvantorifraasi võib olla raskem ära tunda, kui põhjaks on mingi umbmäärast hulka tähistav sõna. Kvantorifraaside leidmiseks kasutatakse ESTKGs kvantorite loetelusid. Nimisõnaliste kvantorite hulka on raske ammendava loeteluna kirja panna, see nõuaks palju tööd leksikoniga. Kitsenduste grammatika leksikonis on praegu enamlevinud nimisõnalised kvantorid: *liiter, meeter, tonn, tund, aasta, tükk, kübe, natuke, ivake, veidike, killuke, korv, viil, kott, ämber, pudel, tass, komplekt, hulk, rühm, polk, trobikond, kamp, punt, kari, kimp, virn, kuhi, osa* jne. (Müürisep 2000: ptk 3.8.5.) Süntaktilisel ühestamisel tekitasid probleeme sellised umbmäärased kvantorid nagu *saadetis, võrdne arv, rohkem, kõige rohkem*. Järgnevalt mõned näited erinevalt analüüsitud lausetest.

Lauses ... *mitte enne, kui saadetis relvi kohal* oli kvantori laiend *relvi* ekslikult analüüsitud nimisõnaliseks järeltäiendiks.

Lauses *Esimeses hääletusvoorus saab vähemalt kaks kandidaati võrdse arvu hääli* on samuti ühestajad olnud eri arvamustel, milline sõnavorm peaks olema fraasi põhi ja milline laiend. Kui fraasi põhi oleks *hääli*, oleks *arvu* nimisõnaline eestäiend, kuid korrektse analüüsina saab *hääli* kvantori laiendi ja *arvu* kvantorifraasi põhjana objekti tõlgenduse.

Veerandi kõikidest kvantorifraasi vigadest tekitas kvantorifraasi süntaktilise funktsiooni määramine – mõnikord on raske otsustada, kas kvantorifraas on lauses adverbiaali või subjekti, samuti ka adverbiaali või objekti funktsioonis. Sihitisekäändelised määrused sarnanevad mõnevõrra sihitisega käändevormistiku ja sellega seonduva aspektivahelduse poolest. Semantiliselt ei ole nad aga tegevusobjektid. Nad ei seondu verbi ega lause transitiivsusega. Nt. *Sõitsime järve äärde kaks tundi. Juuksed kasvavad millimeetri päevas* (EKG II 1993: 49). Samal ajal võib aga kvantorifraas tõepoolest olla ka sihitise rollis, nagu näha lausetest, milles ühestajate arvamused lahknesid.

nendesamade jaoks anti nüüd kaks kuud aega Anti mida või anti kui palju?

Kui sind prostituudiga tabati, võis see tähendada viit aastat sunnitööläagrit ... Tähendab mida? Või tähendab kui palju?

Uute rahatähtede ja müntide juurutamiseks kulub kolm aastat Kulub mis? Või kulub kui palju? Siin on kvantorifraas pigem määruse rollis.

Kas kvantorifraas on aluse või määruse rollis lauses... *saab sellel olema seimis 46 häält* ? Otsustati aluse tõlgenduse kasuks.

Põhiarvsõnad on üldjuhul kvantorifraasi põhjaks, kuid sellest reeglist on mitu erandit.

Mitmeid vigu põhjustas asjaolu, et arvsõna *üks* talitleb atributiivse adjektiivina (EKG II 1993: 142), mitte kvantorifraasi põhjana, nagu teised põhiarvsõnad.

Ka mitmusevormilised (pro)numeraalid (*kümned, sajad, miljardid* jne.) talitlevad atributiivse adjektiivina. Laiendiks on asja või komplekti märkiv nimisõna (EKG II 1993:142).

Nt lauses ... *töötavad Vene sõjaväespetsialistid kabinetivaikuses välja kümneid sarnaseid ettepanekuid* on fraasi põhjaks siiski ettepanekuid, mitte kümneid ja tegemist ei ole kvantorifraasiga.

Kolmandaks erandiks reeglist on *ndik*-lõpulised murdarvsõnad, mis talitlevad nagu osa märkivad kvantornimisõnad (EKG II 1993: 142). Seepärast on järgnevas fraasis järellaiend *aktsiatest* saanud nimisõnalise järeltäiendi tõlgenduse, mitte kvantori laiendi oma.... *viia umbes viiendik telekomi aktsiatest börsile.*

3. Kokkuvõte

Käesolev artikkel kirjeldas käsitsi süntaktiliselt analüüsitud korpuse loomise protsessi. Sellise korpuse koostamise esmane eesmärk on arvutuslingvistiline – korpus on mõeldud eelkõige test- ja treeningmaterjaliks eesti keele automaatse analüüsi süsteemide ja nende rakenduste jaoks. Kuid "lisaväärtusena" annab selline korpus kahtlemata materjali ka lingvistiliseks uurimistööks. Suure tekstihulga käsitsi analüüsil ilmnes ka mitmeid lingvistilisi probleeme, millest mõningaid oleme ka käesolevas artiklis käsitlenud.

Süntaktiliseks märgendamiseks valiti kitsenduste grammatika formalism. Seda põhiliselt seepärast, et tema abil märgendamine sarnaneb kõige enam sellele eesti keele traditsioonilisele grammatilisele märgendamisele, mida filoloogidele õpetatakse: märgendatakse sõnadevahelisi süntaktilisi sõltuvusi, kuid mitte tingimata kõiki (st ei püüta teha täielikku sõltuvuspuud); samas määratakse alati lauseliikmed (mida osad formalismid ei tee). Seetõttu oletasime, et just seetõttu on eesti keele süntaktilise märgendamise traditsiooni kõige lihtsam formaliseerida just kitsenduste grammatikat kasutades.

Tegelikult selgus, et nii lihtne see formaliseerimine ka pole. Paistab, et traditsioonilised grammatikakirjeldused on liiga vähe läbi töötatud või lubavad sama konstruktsiooni erinevaid tõlgendusi. Tulemus: filoloogid ei suuda omavahel kokku leppida, kuidas konkreetseid sõnavorme kontekstis märgendada. Üheks lahenduseks on mitmese lahendi aktsepteerimine, nt määruse ja järeltäiendi puhul. Teiseks võimalikuks strateegiaks on teatud lihtsustamine, eriti nendel juhtudel, kus sõna funktsioon sõltub semantikast või suhteliselt sagedasel konstruktsioonil on võimalikud mitu erinevat tõlgendust.

Kui võrrelda käsitsi süntaktilist märgendamist käsitsi morfoloogilise märgendamisega (vt Kaalep jt 2000), siis on nende puhul rakendatud erinevaid strateegiaid: esimese puhul taotleti lingvistiliselt piinlikult korrektset ühest tulemust, teisel puhul rahulduti aga mitmese ja mõnes lingvistilises aspektis mittekorrektse tulemusega.

Erinevad strateegiad on tingitud sellest, et eesti keele erinevad tasandid on traditsioonilise lingvistika poolt erinevalt läbi töötatud: morfoloogia on rohkem läbi uuritud kui süntaks. Kui morfoloogilisel märgendamisel peab inimene raskel juhul otsustama, milline märgend valida, siis tugineb ta sõna süntaktilisele funktsioonile. Samas olukorras süntaktilise märgendamise puhul tuleb toetuda semantickale. Semantikale tuginemine on aga palju subjektiivsem kui süntaksile, mistõttu objektiivset alust üheseks korrektseks otsustamiseks on raske leida.

Kirjandus

EKG I 1995 = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. Eesti keele grammatika. I. Morfoloogia. Sõnamoodustus. Tallinn: ETA Eesti Keele Instituut.

EKG II 1993 = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. Eesti keele grammatika II. Süntaks. Lisa: Kiri. Tallinn: ETA Keele ja Kirjanduse Instituut.

EKSS 1988–1997 = Eesti kirjakeele seletussõnaraamat. Tallinn: Keele ja Kirjanduse Instituut/Eesti Keele Instituut.

Järvinen, T., Tapanainen, P. 1997. A Dependency Parser for English. Technical Reports, No. TR-1. Department of General Linguistics. University of Helsinki.

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. (toim) 1995. Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text. Berlin/New York: Mouton de Gruyter.

Kaalep, H.-J. 1998. Tekstikorpuse abil loodud eesti keele morfoloogia-analüsaator. – Keel ja Kirjandus 1, 22–29.

Kaalep, H.-J., Vaino, T. 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Käesolevas kogumikus.

Kaalep, H.-J., Muischnek, K., Müürisep, K., Rääbis, A., Habicht, K. 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? – Keel ja Kirjandus, ilmumas.

Müürisep K. 2000. Eesti keele arvutigrammatika: süntaks. Doktoritöö. TÜ arvutiteaduse instituut. Tartu, ilmumas.

Müürisep K. 1998. Eesti keele süntaksianalüsaatorist. – Keel ja Kirjandus 1, 47–56.

Puolakainen, T. 1998. Eesti keele kitsenduste grammatika morfoloogiline ühestaja. – Keel ja Kirjandus 1, 37–46.

Puolakainen, T. 2000. Eesti keele reeglipõhise morfoloogilise ühestamise probleemseid kohti. – Käesolevas kogumikus.

Saagpakk, P. 1992. Eesti–inglise sõnaraamat. Estonian–English Dictionary. Tallinn: Koolibri.

Voutilainen, A., Järvinen, T., 1995. Specifying a shallow grammatical representation for parsing purposes. – Proceedings of the Seventh Conference of the European Chapter of Association for Computational Linguistics. University College Dublin, Ireland. 210–214.

Eesti suulise kõne korpus ja mõnede allkeelte võrdluse katse

Tiit Hennoste, Liina Lindström, Andriela Rääbis, Piret Toomet, Riina Vellerind
Tartu Ülikool

Käesolev artikkel annab lühikese ülevaate suulise kõne uurimisest ning keskendub seejärel suulise kõne korpuste probleemidele. Lõpus võrdleme nelja allkorpuse sadat sagedasemat sõnavormi.[1]

1. Suulise kõne uurimisest

Klassikaline keeleuurimine oli eelkõige kirjalike tekstide keele uurimine. Alates käesoleva sajandi algusest on uuritud ka suulist kõnet. Suur areng toimus selles osas 1960. aastatest, mil tekkis hulk erinevaid uurimissuundi, keda ühendab eelkõige see, et nad uurivad keelt tema kasutuses ja tõlgendavad keelt ja tema kasutust eelkõige funktsionalistlikult kui vahendit millegi jaoks. Nende suundade oluliseks uurimisobjektiks, kuigi erinevatel põhjustel, sai suuline kõne ja suhtlus (vt pikemalt Hennoste 2000b).

Need suunad on uurinud näiteks suulise kõne leksikaalseid ja grammatilisi erinevusi kirjalikest tekstidest ja nende põhjusi. Mõjukaimad uurijad on siin olnud Wallace Chafe ja Douglas Biber, kelle lähenemine on küllalt erinev (Chafe 1982, Biber 1988). Teiseks rühmaks on linnades kõneldava igapäevakeele (argikeele) uurimine ja tema võrdlemine kirjakeele ja vanade kohamurretega. Siia kuuluvad näiteks vene kõnekeele uuringud, mis algasid juba 1960. aastatel ning soomlaste linnakeelte uuringud eelkõige 1970.–80. aastatel (Zemskaja 1979, Suojanen 1985). Teistest erinev lähenemine on konversatsioonianalüüs, mida eelkõige huvitab mitte niivõrd keele kui suhtluse uurimine. Tema eesmärgiks on leida suhtlust juhtivad normid ning nende taustal analüüsida seda, kuidas inimesed igal konkreetsel juhul neid järgivad ning mis juhtub siis, kui nad norme eiravad (vt Tainio 1997; Hutchby, Wooffitt 1998)

---

[1] Käesoleva artikli aluseks olevaid uurimusi on toetanud Eesti Teadusfond (grant nr 3105) ja Teaduskompetentsi nõukogu (sihtfinantseerimise projekt TFLEE 0528)

Suulise kõne uurimiseks on kasutatud mitmesuguseid meetodeid.

Üks meetodite rühm on kvantitatiivne analüüs. Seda lähenemist on kasutanud sotsiolingvistid, registri- ja stiiliuurijad ning keelestatistikud. Meetodi tuumaks on arusaam, et allkeeled erinevad üksteisest eelkõige mitmesuguste keelejoonte kasutussageduse poolest. Mingis allkeeles eelistatakse teatud sõnu ja grammatilisi vorme ning välditakse teisi. Klassikalised keelestatistilised uurimused on tähendanud eelkõige sõnade ja sõnavormide või grammatiliste kategooriate esinemissageduste väljatoomist ning erinevate allkeelte sõnasagedustabelite võrdlemist (vt ülevaadet Tuldava 1977). Registriuurijad on sellele lisanud mitmesugused situatiivsed ja sotsiaalsed taustamõjurid ning proovinud leida püsivaid suhteid keelejoonte sageduse ja taustamõjurite vahel (vt ülevaadet Chafe, Tannen 1987). Douglas Biber tõi võrdlusesse keerukama lähenemise näidates, et erinevad keelelised jooned moodustavad kokkukuuluvaid kimpe ning et erinevad taustategurid on seotud just nende kimpude eelistamise või vältimisega (Biber 1988).

Teine meetodite rühm on kvalitatiivne mikroanalüüs. Selle eesmärgiks pole mitte selle otsimine, mida kasutatakse rohkem või vähem, vaid selle näitamine, mis eesmärgil ja mis põhjusel just selles konkreetses tekstikohas mingit keelevahendit kasutatakse. Sellist suunda esindab väga puhtal kujul konversatsioonianalüüs, aga seda on kasutanud ka interaktiivne sotsiolingvistika ja antropoloogiline lingvistika. Erinevused nende suundade vahel on eelkõige selles, millega seletatakse konkreetsete keelevahendite kasutust antud tekstikohas. Konversatsioonianalüüs kasutab seletamiseks vestluse ülesehitamise universaalseid norme, teised suunad eelkõige mitmesuguseid sotsiaalseid, psühholoogilisi ja situatiivseid tegureid.

Eesti keele uurimine on olnud eelkõige kirjakeele ja murde uurimine. Põhiosa uurimustest on tehtud kirjakeele kohta. Suulise materjaliga on tegelnud murdeuurimused, foneetika ja slängiuurimine. Moodne suulise kõne uurimine sai tõsisema hoo alles aastast 1996, mil Tartu ülikoolis algas Tiit Hennoste iga-aastane suulise kõne loengukursus ning käivitus Eesti linnakeelte kogumise ja uurimise projekt (1997–2000), mille kesksed täitjad on olnud Liina Lindström, Andriela Rääbis, Piret Toomet ja Riina Vellerind. Selle projekti raames on koostatud esimene süstemaatiline suulise kõne korpus, mida ka käesolevas artiklis tutvustatakse ja

analüüsitakse. Selle töö tulemuseks on rida artikleid, magistri- ja bakalaureusetöid. Esimene kokkuvõttev ülevaade eesti suulisest kõnes on praegu ilmumas ajakirjas Akadeemia (vt esimesi peatükke Hennoste 2000b, 2000c, 2000d, 2000e)

2. Transkriptsioon, konsituatsioon ja korpus

Järgnevalt läheme suulise kõne uurimismaterjaliga seotud probleemide juurde. Enne kui saame asuda suulist kõnet uurima on vaja lahendada kolm probleemi. Esiteks, suuline kõne tuleb uurimise jaoks fikseerida. Selleks on tarvis transkriptsiooni. Teiseks on vaja fikseerida need taustamõjurid, mis keelekasutust mõjutavad ehk situatsioon, milles suhtlemine toimub. Kolmandaks, suulise kõne uurimine teeb oma järeldused alati konkreetse keelelise materjali põhjal. Seega on vajalik tekstikorpuse olemasolu.

Siinkohal ei hakka me kirjeldama nende probleemide teoreetilist tausta, mille kohta võib lugeda eraldi artiklist (Hennoste 2000a). Vaatleme vaid neid punkte, mis on olulised TÜ suulise kõne korpuse ja käesoleva artikli väikese kvantitatiivse analüüsi taustaks.

2.1. Suulise kõne transkriptsioon

Pole olemas absoluutset või objektiivset transkriptsiooni. Iga transkriptsioonisüsteem oleneb nendest teoreetilistest alustest, millele uurija toetub ning toob välja kõnes neid aspekte, mis uurijale olulised on. Suulise kõne uurijad on koostanud mitmeid transcriptsioonisüsteeme. Tuntuimad neist on Gail Jeffersoni koostatud konversatsioonianalüüsi transkriptsioon, London–Lundi suulise kõne korpuse transkriptsioon, J. Du Bois' ja tema kolleegide koostatud *Corpus of Spoken American English* transkriptsioon ning lastekeele projekti ja andmebaasi CHILDES transkriptsioon ehk CHAT. CHATi juurde on lisana ehitatud ka konversatsioonianalüüsi transkriptsiooni kasutamise võimalused (vt pikemalt Hennoste 1999; London–Lundi transkriptsiooni kohta Svartvik, Quirk 1980, Peppé 1995; *Corpus of Spoken American English* transkriptsiooni kohta vt Du Bois jt 1993; CHAT on leitav aadressil http://childes.psy.cmu.edu/).

Transkriptsiooni koostamise teoreetilised põhimõtted on pälvinud suhteliselt vähe tähelepanu. Olulisena võib varasemast ajast nimetada ainult lastekeele uurija Elinor Ochsi tööd (Ochs 1979). Hoogu andis transkriptsioonide teoreetilisele analüüsile suulise kõne

arvutiuurimuste levimine 1990. aastatel. Olulisemad teoreetikud on olnud John Du Bois ja Jane Edwards (Du Bois 1991, Du Bois jt 1993; Edwards 1992, 1993a, 1995).

Tavalise transkriptsiooni eesmärke aitab ellu viia kaks rühma printsiipe: kategooria kujundamise (*design*) printsiibid ja loetavuse printsiibid.

Kategooriate kujundamise printsiipidel on tähtsad kolm omadust:

- kategooriad peavad olema süstemaatiliselt eraldusvõimelised, nii et iga keelelise üksikjuhtumi jaoks on alati selge, kas ta on selle kategooria jaoks kohaldatav või mitte. Nt pausi pikkuste klassifitseerimisel peab olema mingi alus, et eristada pikki ja lühikesi pause;
- kategooriad peavad olema ammendavad: iga konkreetse juhtumi jaoks andmetes peab olema kategooria, millesse ta sobib (kas- või "muud");
- kategooriad peavad olema süstemaatiliselt üksteist välistavad: iga juhtumi jaoks sobib ainult üks kategooria. Nt lühikese pausi ülemine piir on pika pausi alumine piir.

2. Loetavuse printsiibid. Kuna tavaliselt on transkriptsioon analüüsitav rida-realt lugemise abil, siis on oluline, et info oleks esitatud vormis, mis lubab uurijal keskse info välja eraldada nii kiiresti kui võimalik. Et eristada eri info suhtelist tähtsust ja eri osade omavaheliste suhete tihedust, kasutatakse kahte printsiipi:

- visuaalne tähtsus (*visual prominence*, nt *bold*, allajoonimine jms);
- ruumiline korrastamine (*spatial arrangement*, nt eri tekstiosade lähedus üksteisele, esitamine vasakult paremale või ülalt alla jne).

Meie kasutatav transkriptsioonisüsteem pärineb konversatsioonianalüüsist (vt pikemalt Hennoste 2000a). See on ette nähtud eelkõige vestluse uurimiseks. Sellest tuleneb ka tema kategooriate valik ning eri kategooriate transkribeerimise sügavus. Selles transcriptsioonis tuuakse välja seitse suurt nähtuste rühma:

- sõnad ja mitmesugused suhtlushäälitused;
- suhtlusüksused;
- pausid ehk mõõdetud intervallid lausungite sees ja nende vahel;
- kõne omadused (intonatsioon, venitused, katkestamised, rõhud, valjus jms)

- pealerääkimised ja haakumised: juhtumid, kus mitu inimest kõnelevad korraga või kus ühe kõneleja jutu lõpp ja teise algus haakuvad tihedalt kokku;
- transkribeerija kahtlused (halvasti kuuldud sõnad vms);
- kirjeldused nähtustest, mille kohta puudub transkriptsioonimärk või mida transkribeerija ei taha transkribeerida, kuid mis on vajalikud ära näidata (hääle omadused, kõrvalised hääled, nutt jms).

Seega on selles transkriptsioonis enam läbi töötatud mittekeelelised nähtused ning kõnevoorude omavaheline seotus dialoogis, st nähtused, mis lubavad sügavamalt analüüsida vestlust kui sotsiaalset fenomeni. Sõnade märgistus on aga oluliselt pealispindsem kui nt foneetilises transkriptsioonis. Selle transkriptsiooni Eesti variandi praegu kasutatavad märgid on kirjas mitmes uurimuses ja kättesaadavad ka internetis (vt Hennoste 2000a, 2000b; http://www.cl.ut.ee). Esitame siinkohal ainult väikese tekstikatkendi konversatsioonianalüüsi transkriptsioonis.[2]

Näitetekst konversatsioonianalüüsi transkriptsioonis

M: kuule mamma niipalju [ei jõuagi ´teha kui nad {-}]
H: [üks üks kaheksakümend kilo ´liha] süia sis ´suure ´perega. (.) oh ´issand. (1.2)
K: ´n:uusata korraks. (0.5)
M: ommegi nii. (3.0) n:ooh? (1.0)
K: @ kook kook @ ((imiteerib kana)) [hehe]
M: @ [no kook] kook, @ [me sööme su {-}]
H: [kas koko ((kukk)) on ´söödud=e.]

2.2. Transkriptsioon ja arvuti

Arvutite kasutamine tõi kaasa mitu transkriptsiooni eriprobleemi. Esimene neist on varasema käsitsi tehtud transkriptsiooni kohandamine arvuti tarvis. Sellega tuli tegelda näiteks esimese suulise kõne arvutikorpuse London–Lundi korpuse loojatel (vt Peppé 1995).

Teiseks esitab arvuti transkribeerimisel omalt poolt mitmesuguseid lisanõudeid. Nende nõuete täitmiseks on loodud arvuti-

[2] Märk osutab rõhulist sõna, : märgib venitust, (.) ja (1.2) pause, ? osutab tõusvat intonatsiooni, = märgib sõnade kokkuhäälduamist, [ ] märgivad samaaegselt kõneldud lausungeid, {-} osutab halvasti kuuldud sõna, @ @ osutavad imiteerimist ia (()) tekst on litereerija kommentaarid.

korpuste omad transkriptsioonisüsteemid. Ülevaateid neist võib leida suulise kõne arvutil uurimise ülevaatekogumikust *Spoken English on Computer* (Leech jt 1995). Samas tekib nende loomisel mitu suurt probleemipundart.

Esiteks, transkriptsioon peab olema kergesti arvutil käsitletav. Arvutianalüüsil on kõige olulisemad kodeerimise süstemaatilisus ja ennustatavus (*predictability*) (Edwards 1993a).

Süstemaatilisus tähendab mittetähendusliku varieerumise vältimist (nt varieerumine häälduses, suurtähed, sõnavahed jms). See toob suulises kõnes kaasa suure probleemi, kuna mitmesugune häälduslik varieerumine on selle keelevormi olemuslik omadus, kusjuures see varieerumine võib olla nii tähenduslik kui ka mittetähenduslik. Vajab pikka uurimist, enne kui võib öelda, kummaga on tegemist. Samal ajal on mittetähenduslik varieerumine tingitud mitmesugustest taustateguritest ja teksti enda omadustest. Kuna üheks oluliseks suulise kõne uurimissuunaks on just selle kindlakstegemine, millised on varieerumise ja taustategurite seosed, siis teeb sellise varieerumise väljajätmine transkribeerimisel suure osa suulise kõne uurimisest arvutil võimatuks.

Ennustatavus tähendab süstemaatilisuste kasutamist, mida uurijad saavad pakkuda oletamiseks, kuidas neid huvitav vorm on eeldatavasti kodeeritud enne kui täpsustada otsimiskäsku nende leidmiseks. Kui uurijad teavad ainult ühte varianti, kuigi korpuses on neid palju, siis arvuti otsib ainult ühe osa asjassepuutuvatest nähtustest ja tulemuseks on suured vead analüüsis. See probleem on väga väike kirjakeele puhul, kus on tegu väheste ning sealjuures norminguliste ja selliselt küllalt täpselt ennustatavate variantidega. Kuid suulise kõne puhul pole selline süstemaatilisus saavutatav, sest tekstides on samast sõnast alati mitmeid variante, lisaks mitmeid *ad hoc* keelendeid. Ja variante saab ennustada veidigi täpsemalt alles peale seda, kui on uuritud läbi väga suur hulk tekste. See nõuab omakorda arvutianalüüsi ja nii tekib omamoodi surnud ring.

Teiseks tekib siin probleem selle tõttu, et transkriptsiooni arvutil kergesti käsitsetavuse printsiibid ja loetavuse printsiibid on omavahel väga halvasti sobitatavad.

Transkriptsioonid on varem olnud kasutatavad uurija enese poolt ilma arvuti abita, st nad on tehtud võimalikult uurijasõbralikud. See tähendab nii seda, et transkribeerimine on uurijatele lihtne kui ka seda, et selle transkriptsiooni kasutamine on lihtne. Selliselt

transkribeeritud tekste saab tänapäeval ka suhteliselt lihtsalt arvutil statistiliselt uurida (otsida vajalikke tekstikatkeid, teha erinevaid sõnaloendeid ja statistikat).

Sellise transkriptsiooni näitena esitame siinkohal CLAN-i.

CLAN on programm, mis on loodud selleks, et analüüsida andmeid, mis on kirja andud CHAT (*Codes for Human Analysis of Transcripts*) transkriptsioonis. CHAT on algeselt loodud lastekeele uurimiseks kuid kohandatud hiljem ka vestlusanalüüsi transkribeerimiseks Jeffersoni transkriptsiooni märke kasutades. CHAT transkriptsioon on kasutatav nii lugemiseks kui ka arvuti tarvis, st puudub vajadus kasutada kahte transkriptsiooni. CLAN lubab praegu analüüsida näiteks kollokatsioone, sõnasagedusi (sh erinevates lausepositsioonides), leiab pikimaid sõnu, arvutab vormide ja lausungite keskmisi pikkusi, mõõdab foneemide sagedusi jms (CLAN pdf, 37–128). CLAN on avatud programm, st temasse tehakse pidevalt uusi võimalusi juurde.

CLANi miinuseks on see, et tal pole midagi pistmist kirjalike tekstide märgendamisega. See ei luba kasutada sama programmvarustust.

Näitetekst CLAN-i versioonis

*MOT: kuule mamma niipalju <ei jõuagi teha [!] kui nad xx> [>] +/.
*GRM: <üks üks kaheksaküm(m)end kilo liha [!]> [<] süia sis suure [!] perega [!].
*GRM: oh issand [!].
*DAU: nuusata [!] korraks.
*MOT: ommegi nii.
*MOT: n:ooh?
*DAU: <kook kook> [% imiteerib kana häälitsust].
%par: DAU naerab
*MOT: no kook kook, <me sööme su xx> [>].
*GRM: <kas koko [: kukk] on söödud [!] e> [<]?

Kui aga soovitakse teha arvuti abil keerukamaid analüüse, siis on tarvis korpus transkribeerida maksimaalselt arvutisõbralikult. Tulemuseks võib olla see, et selline transkriptsioon ei ole enam inimesele loetav ning korpus peab olema transkribeeritud kahel viisil, inimese jaoks ja arvuti jaoks. Selle probleemi üle diskuteerivad pikalt tippasjatundjad raamatus *Spoken English on Computer* (Leech jt 1995).

Arvutitranksriptsiooni näiteks on seal valitud praegu tuntuim programm TEI.

TEI on uurimisprojekt, mille eesmärk on välja töötada täielik komplekt juhiseid, mis võimaldavad ükskõik millise eriala uurijatel viia tekste arvuti poolt loetavasse vormi sõltumata sellest, millise

riistvara või tarkvaraga on tegemist ja sõltumata keelest. TEI on eelkõige mõeldud ja kohandatud kirjalike tekstide märgistamiseks. Siiski on tal olemas ka lisasoovitused suulise kõne märgendamiseks (vt Johansson 1995: 82–98).

TEI formaadis tekst koosneb kahest osast: teksti pea (*Header*) ja tekst ise. Pea dokumenteerib lisainfot iga teksti juurde (arvutifaili bibliograafline info, teksti loomise kohta käiv info, mittebibliograafiline info teksti kohta (nt osalised, olukord jms) ning elektroonilise teksti tegemise ajalugu). See osa on suulise ja kirjaliku teksti puhul praktiliselt sama, kuigi suurt osa suulise suhtluse uurimiseks vajalikust taustainfost ei ole võimalik paigutada TEI poolt pakutud *Header*i lahtritesse.

Teine osa on tekst ise. Teksti osade märgistuses ei tee TEI vahet erinevat tüüpi info vahel, esitades sama tüüpi märkidega nii teksti struktuurielemente kui ka erinevat interpretatiivset infot. TEI märgenduse põhimõisteks on märgend, mis koosneb nurksulgudes koodist, näiteks *<name>*. Märgid paiknevad tavaliselt (kuid mitte alati) nii vastava üksuse alguses kui lõpus. Põhimärgenditele on võimalik lisada mitmesuguseid atrbuute. TEI märgendidi moodustavad hierarhilise süsteemi. Näiteks suulise teksti puhul soovitatakse selliseid märke.

- tekst *<text>*;
- teksti allosa (*subdivison *), mis sisaldab mitut lausungit, mida peetakse vajalikuks koos käsitelda;
- lausung (*utterance*, *<u>*), st kõnelõik, mis on tihti piiratud pauside või kõnelejate vahetusega. Tegu on tegelikult formaalse kõnelõiguga;
- *<s>* lausungi allosa, mis on välja toodud süntaktiliste või prosoodiliste kriteeriumide alusel.

Lisaks tuuakse välja suulise kõne erinähtused, mis on esitatud järgmise formaalse loogilise skeemina:

- kommunikatiivsed vokaalsed, kuid mitte ilmtingimata leksikaalsed nähtused (pausitäitjad ehk üneemid, mitteleksikaalsed tagasisided jms) *<vocal>*;
- kommunikatiivsed nähtused, mis ei ole vokaalsed (kehaliikumised jms) *<kinesic>*;
- muud mittevokaalsed ja mittekommunikatiivsed nähtused, mis toimuvad suhtluse ajal ja võivad seda mõjutada *<event>*;
- kirjutatud tekst suulise teksti sees *<writing>*;

- pausid *<pause>*;
- muutused hääle kvaliteedis *<schift>*.

Näitetekst TEI versioonis
(parema loetavuse huvides on täpitähed jäetud SGML koodi teisendamata)

```
<u who=M> <seg>kuule mamma niipalju <anchor synch=T1> ei jõuagi teha
kui
nad (-) <anchor synch=T2> </seg> </u>
<u who=H> <seg> <anchor synch=T1> üks üks kaheksakümend kilo liha
<anchor
synch=T2> süia siis suure perega. </seg> <pause dur=0.2> <seg> oh issand
</seg>
</u> <pause dur=1.2>
<u who=K> <seg> nuusata korraks </seg> </u> <pause dur=0.5>
<u who=M> <seg> ommegi nii </seg> <pause dur=3.0> n:ooh? </u> <pause
dur=1.0>
<u who=K> <seg> @ kook kook @ ((imiteerib kana)) <anchor synch=T3>
hehe
<anchor synch=T4> </seg> </u>
<u who=M> <seg> <anchor synch=T3> @ no kook <anchor synch=T4>
kook, @
<anchor synch=T5> me sööme su <anchor synch=T6> </seg> </u>
<u who=H> <seg> <anchor synch=T5> kas koko ((kukk)) on söödud=e.
<anchor
synch=T6> </seg> </u>
```

Selles lõigus on praegu märkimata rõhud, imitatsioon, venitused ja sõnade kokkuhäälduses, samuti ebaselgelt öeldud sõnad. Selle põhjuseks on eelkõige TEI nõrgad kohad.

TEId on paljud tugevalt kritiseerinud (nt Sinclair 1995). Payne (vt Johansson 1995: 95–97) on leidnud, et üldiselt on arvutisõbralikud TEI skeemid otse seotavad erinevate kasutajasõbralike kodeerimissüsteemidega. Lisaks toob ta aga välja hulga probleemseid kohti. Olulisemad probleemid TEI rakendamisel suulisele kõnele on järgmised.

- Ta on inimesele praktiliselt loetamatu, eriti siis kui tegu on suulise tekstiga, kus on igasuguseid lisamärke. Lisaks ei luba TEI kasutada täpitähti, mis teeb suure osa tekstist eesti keeles praktiliselt loetamatuks. See tähendab, et nende analüüside jaoks, mida arvutiga teha ei saa või ei taheta, on tarvis sama tekst transkribeerida teise transkriptsiooni. Seega on vaja tarkvara, mis muudaks tavalise lugejasõbraliku transkriptsiooni TEIks või vastupidi. Viimane pole aga eriti mõttekas, sest nõuab eraldi õppinud inimesi transkribeerima ning lisaks on suuline suhtlus selline, mis vajab esialgu väga palju tavalist analüüsi, et jõuda arvutianalüüsiks piisava tasemeni. Sellist

tarkvara, mis muudaks tavalised tranksriptsioonid TEI formaati, seni teada ei ole.

- TEI annab liiga primitiivse ja jäiga skeemi suulise teksti liigendamiseks lausungiteks ning tema defineeritud lausung ei sobi tegeliku lausungiga kokku. Nt TEI ei luba lausungit katkestada või välja tuua kahe inimese poolt koos loodud lausungit, kuid see on tavaline nähtus suulises vestluses (vt Sinclair 1995: 108). Suur probleem on sellega, et suulises kõnes ei ole üksuste vahel aredaid piire nagu kirjas punktid lausete lõpus. Suuline kõne liigendab teksti paralleeelselt süntaktiliselt, semantiliselt ja intonatsiooniliselt, kuid need piirid ei pruugi konkreetses tekstikohas kattuda (vt Hennoste 2000b). See on lahendamatu probleem TEI jaoks, sest TEI ei luba kasutada näiteks *<s>* märki korraga nii prosoodia kui süntaktilise liigenduse jaoks. Uurija peab tegema valiku, kuid uurimise jaoks tähendab see paljudel juhtudel lihtsalt vale liigendust.
- TEI ei luba kasutada teatud märke (näiteks allajoonimist), mis on tavalised tranksriptsioonides eelkõige rõhuliste sõnade ja sõnaosade märkimiseks.
- TEIs on väga keeruline tulla toime suulise suhtluse loomuliku nähtuse, nimelt korraga rääkimisega. TEI on proovinud seda lahendada järgmiselt. On võetud kasutusele tühi märk *<ancor>*, mida saab täita erineva sisuga. Pealerääkimise märkimiseks listakse sinna atribuut *synch* ja selle number (*<ancor synch=T1>*). Selle juurde kuulub eraldi ajarida (*timeLine*), mis osutab, millised pealerääkimised kuuluvad kokku. timeLine on sealjuures mõeldud laiemalt igasuguste elementide ajaliseks sünkroniseerimiseks (Johansson 1995: 91–93). Selline põhimõtteline lisandus TEIle nõuab aga lisaks omaette tarkvara.
- Suulises kõnes on väga oluline prosoodia. See on keerukas võrgustik, mida võimalikult põhjalikult paneb kirja foneetiline transkriptsioon ja mida muud süsteemid proovivad lihtsustatult kirja panna. TEI pole sellele erilist tähelepanu pööranud ja seetõttu on temas väga keerukas ja kirja panna näiteks lauserõhke, intonatsioonimuutusi jms. Ka neid soovitatakse panna eraldi reale.
- TEI kasutamine oleks kasulik, sest see võimaldaks analüüsida sama keele suulist ja kirjalikku kõnet sama programmvarustusega (nt kasutada sama süntaksi- või morfanalüsaatorit). Kuid

kahjuks ei ole saadud tulemused näiteks süntaksis võrreldavad. Suulise kõne liigendub põhimõtteliselt teisiti kui kirjalik tekst, mis tähendab, et me võrdleme tegelikult erinevaid üksusi. Kui aga otsida kõnest välja kirjaga analoogilised süntaktilised üksused ning piiritleda need transkriptsiooni abil, siis saaksime üsna absurdse liigenduse, milles mõnedele kõnelõikudele ei ole üldse kohta. Võrrelda saaks üksnes morfoloogiat ja sõnavara, mis aga ei vaja TEI abi.

2.3. Suulise kõne korpus

Tekstide arvutikorpused ja neil põhinev korpuslingvistika moodustab tänapäeval eraldi keeleteaduse haru, mida me siinkohal täpsemalt ei kirjelda (vt McEnery, Wilson 1997; Muischnek 1998). Tuleb vaid nentida, et valdav osa korpusi on siiani olnud kirjutatud keele korpused ja et korpuse tegemise ja arvutianalüüsi meetodid on välja töötatud just kirjalike korpuste peal.

Suulise kõne uurimine on algusest peale olnud tekstidest lähtuv uurimine. See on tähendanud alati ka vajaliku kõnekogu olemasolu. Varasemad suulise kõne materjalid olid lihtsalt kuulmise järgi üles märgitud lause- või sõnatranskriptsioonide sedelkogud.

Lindistatud suulise kõne kogusid hakati tõsiselt korjama alates 1960. aastatest. Näiteks vene uurijad E. A. Zemskaja juhtimisel kogusid vene argikeelt Moskvas ja Leningradis (need tekstid on ilmunud ka kogumikuna (Russkaja 1978)). Soomes koguti suulist kõnet linnakeelte uurimise tarvis 1970. aastatel (Suojanen 1985). 1985. aastal alustas tekstide kogumist Auli Hakulise konversatsioonianalüüsi töörühm Helsingi ülikooli juures.

Suulise kõne arvutikogud võime laias laastus jagada kahte rühma. Üks osa neist on arvutis olevad tekstikogud. St nad on kogutud enam või vähem süstemaatiliselt, kuid on mingis tavalises formaadis (Word, Word Perfect, txt) olevad transkribeeritud tekstide kollektsioonid arvutis. Teise rühma moodustavad tekstikorpused, mis on lisaks ka arvutil töödeldavad ja uuritavad. Osa neist on planeeritudki arvutikorpustena, osa aga on esialgu olnud tekstikogud, mis hiljem on viidud korpuse formaati. Sealjuures on varasemad korpused ca 0,5–1 miljoni sõnalised, viimase aja inglise keele korpused aga sisaldavad juba kümneid miljoneid sõnu.

Suulise kõne korpuseid hakati tegema alguses kirjaliku korpuse ühe osana. Esimene oluline suulise kõne korpus on *London–Lund Corpus of Spoken English* (LLC), mida koguti 1960–70. aastatel Jan Svartviki juhtimisel ja mis sisaldab 500 000 sõna, mis on valitud 100 tekstist. Korpus sisaldab nii monolooge kui dialooge. Monoloogid jagunevad spontaanseteks ja ettevalmistatuteks, dialoogid on jagatud vestlusteks ja avalikeks diskussioonideks (vt Svartvik 1990).

1980. aastatest alates on koostatud palju erinevaid suulise kõne korpuseid, eriti inglise keele kohta (vt Leech, Myers 1995; Edwards 1993b). Sealjuures on tegu suurte projektidega, mille üheks osapooleks on tüüpiliselt mõni ülikool ja teiseks osapooleks erinevad sõnaraamatukirjastused.

Tuntuimad inglise suulise keele korpusi on kolm. Üks on COBUILDi kõnekorpus, mida on tehtud Birminghami ülikoolis John Sinclairi juhtimisel ja mille suurus on kümneid miljoneid sõnu. See on osa ülisuurest monitorkorpusest *Bank of English*. 1990 algas projekt *Corpus of English* (ICE), mida koordineerib Sidney Greenbaum ja mis kogub eri maades kõneldavaid inglise keele variante. Selle ühe osana oli mõeldud ka Ameerika suulise kõne korpus (CSAE), mida veab Santa Barbara ülikool Californias. Kolmas suur korpus on *British National Corpus* (BNC), mille suurus umbes 10 miljonit sõna. Selle alusena rõhutakse kõnelejate sotsiaalsete ja territoriaalsete karakteristikute arvestamist ehk demograafilist mudelit ning see jaguneb nelja võrdsesse rühma: haridustekstid (loengud, koolitunnid jms), äritekstid (ärikõnelused, konsultatsioonid jms), avalikud institutsionaalsed tekstid (poliitikute kõned, jutlused jms) ning vaba aja tekstid (lõunalauavestlused, telefonikõned jms) (vt Crowdy 1993). Viimastel aastate on üha enam alustatud ka muude keelte arvutikorpusi (hollandi, portugali, itaalia, sloveeni jms keeled).

Varasemad korpused olid ja on eelkõige keeleuurimise korpused, mida kasutati ka praktilistel eesmärkidel, nt sõnastike tegemiseks. Viimasel ajal on suulise kõne materjali hakatud koguma ja transkribeerima ka kõnesünteesi ja kõneanalüüsi huvides, st nende kasutamise eesmärgid on arvutuslingvistilised. Sellised andmebaasid on tavaliselt väga suured, kuid nende transkriptsioonid enamasti väga pinnapealsed. Praktiliselt tähendab nende kirjapanemine üksnes sõnade fikseerimist tavalise ortograafia abil. Vaid paarsada tuhat sõna transkribeeritakse tavaliselt sügavamalt. Andmebaase on

koostamisel nii üksikute keelte kohta kui ka mitmele keelele paralleelselt. Osa neist on universaalsed, st haaravad erinevaid allkeeli, kuid üha enam tehakse kitsaid andmebaase mingi konkreetse valdkonna või ülesande tarvis (nt autoteenindus, infotelefon jms). Sellised projektid on riiklikult või Euroopa Liidu poolt finantseeritavad ning nendega tegelevad korraga mitmed uurimiskeskused ja ülikoolid, kes kasutavad lindistuseks ja litereerimiseks lisaks vabatahtlike abi (vt ülevaateid SICLRE II 2000).

2.4. Tartu ülikooli suulise eesti keele korpus

Eestis on tekstikorpusi tehtud alates 1980. aastate lõpust Tartu ülikoolis ja Eesti Keele Instituudis. TÜ eesti filoloogia osakonnas on alates 1980. aastate lõpust tehtud avaliku kirjaliku keelekasutuse ehk kirjakeele korpuseid (Hennoste 1996; Hennoste jt 1998; Hennoste, Muischnek 2000).

Suulise kõne materjale on Eestis kogutud juba 1920. aastatest. Need olid alguses murdematerjalide sedelid, kuhu oli kuulmise järgi märgitud sõnu ja lauseid. Hiljem on kogutud murdetekse lindistuste abil. Nende lindistuste põhiprobleemiks on, et tegu on praktiliselt ainult intervjuudega, kus lühikeste küsimuste abil püüti meelitada informant kõnelema. Seega pole tegu loomulikus situatsioonis kõneldud tekstidega. Väike osa neist lindistustest on transkribeeritud soome-ugri foneetilisse transkriptsiooni ja trükitud. Sealjuures on tekstide trükivariandid tavaliselt tugevalt puhastatud, st neid on lühendatud, visatud välja kõikvõimalikke takerdusi jms. Praktiliselt tähendab see, et nende põhjal ei ole võimalik uurida näiteks tegelikku suulise kõne süntaksit ning ka selliste morfoloogiliste vormide kasutamist, mille varieerumine oleneb sõnavormi asukohast ja rollist lauses, nagu nt *nud*-kesksõna. Lisaks on kogutud slängisõnu väljaspool konkreetset konteksti.

Suulise kõne korpust on tehtud alates 1997 aastast Tiit Hennoste juhtimisel. Lindistused on teinud suulise kõne töörühm (Liina Lindström, Andriela Rääbis, Piret Toomet, Riina Vellerind) ning suulise kõne loengusarjade kuulajad Tartu ülikoolis.

See korpus on planeeritud avatud korpusena, st ta piirsuurust ei ole määratud. Temasse on mõeldud koguda erinevat tüüpi suulist kõnet, nii argisuhtluse kui avaliku suhtluse keelekasutust, nii spontaanset kui ettevalmistatud kõnet, nii monolooge kui dialooge.

Käesolev korpus on koostatud selliselt, et tegu oleks maksimaalselt autentsete situatsioonidega, mida magnetofon ja lindistaja võimalikult vähe mõjutavad. Seetõttu on eelistatud salajasi lindistusi ning selliseid situatsioone, mis on kõnelejatele loomulikud.

Korpuse litereerimisel kasutatakse konversatsioonianalüüsi transkriptsiooni. Transkriptsioonides on kõik nimed ja identifitseerimist võimaldavad numbrid muudetud. Igale lindile on lisatud taustakirjeldus, milles on iga konkreetse lindi puhul lisatud niipalju taustaandmeid kui võimalik (taustakirjelduse skeemi kohta vt Hennoste 2000a; taustakirjelduse lühiskeem on kättesaadav ka internetis).

Uurimisrühma lindikogusse kuulub mitmesuguseid linte. Argisituatsioonide lindid sisaldavad tüüpiliselt umbes pooletunnilist argivestluste lõiku. Ka avaliku suhtluse lindid on 20–30 minutit pikad, kuid neil võib olla mitu erinevat situatsiooni (nt rida kauplusedialoogе). Lisaks on olemas videokassetid.

Osa linte on litereeritud ning arvutisse viidud. Argisuhtluse ja pikkade ametlike suhtluste litereeritud tekstikatked on tüüpiliselt umbes 5 minuti pikkused. Lühemad ametivestlused ja kõik telefonivestlused on litereeritud tervikuna. 2000. aasta maikuu seisuga oli korpuses 259 linti. Korralikult litereeritud ja arvutisse viidud on tekste 182 lindilt, kokku 386 teksti või tekstikatket. Litereeritud tekstide kogupikkus on 230 824 tekstisõna. Lisaks on korpuses 12 videokassetti, millest on litereeritud 4 telesaate katkendit.

Litereerimata või ebatäpselt litereeritud on umbes 80 mitmesugust-silmast silma vestlust või telefonikõnet ning umbes 15 TV ja raadiosaadet. Litereerimata on ka kogu lindistusi, mis tehti Riigikogu valimiste eelsetest tele- ja raadiosaadetest, kokku 22 kassetti ja 29 videolinti.

See korpus on liigendatav mitme parameetri järgi. Suuline kõne ei ole ühtne vaid tema sees on mitmeid erinevaid allkeeli. Allkeeled jagatakse sotsiolingvistikas tavaliselt kahte suurde rühma: kasutajakeskselt defineeritud murded ning situatsioonikeskselt defineeritud registrid.

Murded jagatakse kohamurreteks/dialektideks ja sotsiaalmurreteks/sotsiolektideks. Sotsiolingvistika on leidnud mitmesuguseid seoseid inimese sotsiaalsete parameetrite ja tema keelekasutuse vahel. Sealjuures on erinevad uurijad toonud välja põhiosas samad mõjurid, millest eesti kultuuris on olulised inimese sotsiaalne

päritolu ja staatus (klass, kiht), sotsiaalne võrk (naabrus jms), sugu, vanus ja eriti haridus. Neid parameetreid pole siiani korpuse korjamisel arvesse võetud. Võib vaid öelda, et praegu on korpuses enam naisi, haritud inimesi ja noori või keskealisi inimesi.

Eri murdealadelt pärit inimesed ei kõnele tänapäeval enam enamasti murret, kuid nende keelekasutuses on säilinud mitmesuguseid murdejooni. Samal ajal on murdelisus linnades väiksem kui maakohtades. See korpus on linnakeele korpus, st tema kõnelejad on pärit linnadest. Et võtta arvesse võimalikku erinevat murdetausta, on korpusesse võetud kõnelejaid valdavalt kolmest suurest linnast: Tallinnast, Tartust ja Pärnust. Igal linnal on erinev murdetaust: Tartul Tartu ja Võru murre, Pärnul läänemurre, Tallinnal nii läänemurre kui ka keskmurre. Lisaks on nt Karl Pajusalu seostanud just Tallinnamaa keelega eesti madala rahvakeele mõningaid erijooni (Pajusalu 1997). Praegu on korpuses Pärnu ja Tallinna inimesi siiski vähem kui Tartu inimesi.

Teine suur rühm allkeeli on registrid. Siin on eri uurijad välja toonud erinevaid joonteloendeid ning nimetanud saadud komplekte erinevalt (vt ka Hennoste 2000a, 2000b). Kõige enam keelt mõjutavad situatiivsed parameetrid on:

- suhtlusviis ehk meediumi omadused: kõne/kiri, dialoog/monoloog, spontaansus/redigeeritus;
- füüsilised tingimused, suhtlusolukord (vahetu/vahendatud, argine/ametlik).

Üks meediumipiir on suulisus ja kirjalikkus ise, millest meil siin juttu ei tule teha. Teine oluline meediumijaotus on spontaansus ja redigeeritus, mis eristab situatsioone, mis võimaldavad teksti täpsemat redigeerimist situatsioonidest, kus selline redigeerimine pole võimalik. Just seda piiri on Douglas Biber näinud kõige tugevamana inglise tekstide registriliigenduses. Samal ajal kattub see piir väga suures osas suulise/kirjaliku piiriga, sest suulise teksti tegemine on alati vähemalt 10–15 korda kiirem kui kirjaliku tegemine (120–180 sõna minutis *contra* 10–15 sõna minutis). See aga tähendab praktiliselt, et suuline kõne on alati oluliselt spontaansem. Seetõttu me selle parameetri järgi suulisi tekste korpuses ei liigenda.

Teine spontaansusega seotud parameeter on tavaliselt olnud piir peast esitatavate ja paberilt mahaloetavate või päheõpitud tekstide vahel. Viimasesse rühma kuuluvad nt raadio- ja teleuudised, näidendiesitused jms. Meie korpus sisaldab praegu ainult peast

esitatavaid tekste. Ka korpuse massimeedia tekstid on sellised, kus saatejuht ja külalised vestlevad omavahel küll varem kokkulepitud ja osalt ka läbiarutatud teemal, kuid ei esita varem päheõpitud juttu.

Kolmas meediumipiiripaar on dialoog/monoloog. Klassikaliseks monoloogiks peetakse juhtu, kus üks partneritest arendab juttu ja teisel puudub õigus või võimalus pakkuda vestluse edasiviimiseks omapoolset teksti (loengud, ettekanded, lugude jutustamine). Kuulaja annab siin kõnelejale vaid tagasisidet (noogutused, *mhmh*, naer, tukkumine jms).

Kuid eelkõige konversatsioonianalüüs on näidanud, et tagasisidevõtted annavad alati ka juhiseid selle kohta, kuidas kõneleja peaks jätkama ning kõneleja modifitseerib oma edasist juttu pidevalt vastavalt sellele. Seega on ka sellise teksti puhul rangelt võttes tegu dialoogiga. Siiski, keeleliselt erineb selline tekst selgelt lühikesest repliikidevahetusest klassikalises dialoogis. Keerukaks teeb asja aga see, et eriti argisituatsioonides vaheldub repliikidevahetus pidevalt pikemate lugudega. Kas need pikemad lood kuuluvad ühte rühma klassikaliselt monoloogideks peetavate tekstidega või mitte, on lahtine küsimus.

Käesolev korpus on ülivaldavalt dialoogikorpus kitsas mõttes. Nende hulka kuuluvad kõik argivestlused ja telefonivestlused ning valdav osa silmast-silma ametlikest vestlustest. Puhtad monoloogid on mõned loengud ja jutlused. Pikkadest monoloogilõikudest koosnevad tekstid on eelkõige intervjuud, koosolekud ja koolitunnid.

Neljas oluline piir on vahetu suhtluse ja vahendatud suhtluse vahel. Vahendatuse all mõeldakse kahte eri asja. Üks lähenemine nimetab vahendatuks kõiki suhtlusi, kus kasutakse mingit tehnilist vahendajat, nt telefoni vms. Teine lähenemine tõmbab piiri selle põhjal, kas on või ei ole võimalik vahetu tagasiside. Sel juhul on telefonivestlused vahetud vestlused, sest seal on võimalik anda kohe tagasisidet. Vahendatud oleks sellised suulised tekstid, mis seostuvad massikommunikatsiooniga ning erinevad esimestest ka selle poolest, et vastuvõtja on saatjale anonüümne hulk inimesi. Mõlemat parameetrit arvesse võttes saame tekstid jagada kolme rühma: massimeediatekstid, telefonivestlused ja silmast-silma vestlused. Silmast-silma vestlusi on litereeritud 221 katket, telefonivestlusi 145 (tervikuna litereeritud) ja massimeediatekste 20.

Viimane oluline piir on argine/avalik suhtlus. See piir ei toetu ühele parameetrile nagu eelnevad piirid, vaid on ise kompleksne

süsteem, mille all on inimestevahelised suhted (tuttavad, sõbrad, lähedased *vs* võõrad), inimeste rollid vastavas situatsioonis (eraisik *vs* institutsiooni esindaja), konkreetne füüsiline olukord (kodus, kohvikus jms *vs* tööl, ametiasutuses jms), suhtluse põhieesmärk (osalemine, enese sidumine suhtlusega *vs* informatiivne olemine). Lisaks sellele on argises situatsioonis tavaliselt vooruvahetus vaba, inimeste rollid võivad muutuda sama vestluse jooksul. Avalik situatsioon on rangete rollidega ja palju rangema ülesehitusega. Seal on määratud, kes kõneleb ja mida, kes juhib ja kes allub, kes küsib ja kes vastab, mis järjekorras kõneldakse jms (näiteks arsti–patsiendi vestlus või poedialoog).

Neid tunnuseid kombineerides saame kahe selge tuuma ja hajuvate piiridega kontiinuumi, mille üheks keskmeks on puhas argivestlus ja teiseks puhas institutsionaalne suhtlus.

Korpuses kuulub argivestluste tuuma 76 silmast-silma argivestlust, lisaks 25 eratelefonikõnet. Neis on tegu tuttavate või lähedaste inimestega, kes vestlevad kui eraisikud kodus või muus mitteinstitutsionaalses kohas ja kus suhtlemise oluline eesmärk on vestlemine ise, st vesteldakse igasugustest asjadest, mis pähe tuleb.

Perifeersema, kuid selgepiirilise argivestluste rühma moodustavad 9 telefonikõnet, kus helistatakse tuttavale töö juurde või helistab inimene töö juurest tuttavale koju, kuid aetakse argiasja.

Avalikku poolde kuulub kõige suuremana tuumrühm, kus toimub silmast-silma vestlus, suhtlejad on võõrad, neist vähemalt üks on selles suhtluses mingi institutsiooni esindaja (müüja, arst jms), suhtlus toimub avalikus või institutsionaalses kohas ja tema eesmärk on selgelt informatsiooniline. Siia kuuluvad massimeediasaated, suhtlused mitmesugustes teenindus- ja ametiasutustes (kauplused, postkontor, jaama kassa, raamatukogu, muuseum, reisibüroo, juuksur, polikliinik, maksuamet, haigekassa jms). Teiseks kuuluvad siia loengud, jutlused ja mõned intervjuud. Sealjuures võib neis situatsioonides olla ka lõike, kus on tegu mingi muud tüüpi suhtlusega (nt abikaasadest ostjad räägivad poes mitte ainult müüjaga vaid ka omavahel). Selliseid situatsioone on litereeritud üle saja. Põhiosa neist moodustavad dialoogid kaupluses ja mitmesugustes ametiasutustes.

Telefonivestluste analoogilise tuumrühma moodustavad kõned, kus eraisik on helistanud ametiasutusse ametiasjus. Vaid 6 kõnet on sellised, kus vestlevad omavahel ametiisikud. Neile lisandub 47

telefonimüügivestlust, mis kuuluvad samasse rühma ning moodustavad korpuses omaette sarja.

Avaliku suhtluse teise rühma moodustavad suhtlused, milles suhtlejad on omavahel tuttavad, kuid ajavad ametiasja, st vähemalt üks neist on mingi institutsiooni esindaja ja suhtlus toimub enamasti institutsionaalses kohas (harva ka mujal: ühe poole kodus, mingil neutraalsel pinnal, näiteks õues või tänaval). Siia kuuluvad koolitunnid, korteriühistu koosolek, üliõpilase ja õppejõu suhtlus, õpetaja ja lapsevanema suhtlus, ka mõned arsti ja patsiendi vestlused ja intervjuud). Selliseid situatsioone on korpuses litereerituna umbes nelikümmend. Sellesse rühma kuuluvad vaid mõned telefonikõned.

Kolmanda avaliku suhtluse rühma moodustavad suhtlused, kus inimesed on võõrad, kuid nad tegutsevad eraisikutena neutraalsel pinnal ja hangivad infot. Sellised on näiteks teeküsimised tänaval.

Ülejäänud situatsioonid esindavad mitmesuguseid üleminekurühmi.

Tabel 1. TÜ suulise kõne korpuse struktuur.
Litereeritud tekstid mais 2000

SILMAST-SILMA VESTLUSED (221 vestlust)
- 76 ARGIVESTLUST
- 145 AMETIVESTLUST

Tuumrühm (võõrad suhtlejad)
- 52 kaubandusdialoogi
- 18 teenindusdialoogi (postkontor, jaama kassa, kellaparandus, raamatukogu, kingaparandus, paljundus, juuksur, reisibüroo jm)
- 7 muud vestlust ametiasutustes (muuseumis, registratuuris jm)
- 4 loengut
- 5 intervjuud

Tuttavad suhtlejad
- 14 koolisuhtlust (õpetaja-lapsevanema vestlus, koolitund, bakalaureusetöö kaitsmine, üliõpilase-õppejõu vaheline vestlus)
- 4 majanaabrite vahelist vestlust
- 4 suguvõsa kokkuleku vestlust
- 6 arsti-patsiendi vestlust
- 9 muud dialoogi (raha laenamine, jutlused, koosolekud, intervjuud)

Muud
- 19 võõraste vestlust tänaval
- 3 muud (aktiivne müük, debatt, turu-uuring)

TELEFONIVESTLUSED (145 vestlust)
- 33 ERAVESTLUST (25 tuumrühma vestlust, 9 töö juurde helistamist)
- 110 AMETLIKKU KÕNET
- 2 VALEÜHENDUST

RAADIO- JA TV-SAATED 20

A. Rääbise andmetel on litereeritud tekstikatkete keskmine pikkus 535 tekstisõna, minimaalselt 16 ja maksimaalselt 2561 sõna. Siin-

juures tuleb arvestada, et korpusesse on valitud pikematest vestlustest umbes 5 minuti pikkused lõigud ja lühemad vestlused tervikuna. Argivestlused on üldjuhul valitud lõigud pikemast vestlusest, pikkusega 182–2561, keskmise pikkusega 912 tekstisõna. Avalikud vestlused on enamasti terviklikud situatsioonid, pikkusega 16–2101 sõna, keskmise pikkusega 445 sõna.

Telefonikõned on terviklikud vestlused ning on keskmiselt lühemad kui silmast-silma vestlused. Nende pikkused on 16–1735 sõna, keskmine 313 sõna, sealhulgas erakõned 28–1735, keskmiselt 529 sõna ja ametikõned 22–837 sõna, keskmiselt 220 sõna.

Korpus on esialgu Wordi ja txt formaadis arvutitekstikogu, mis on kasutatav arvutis olevate või paberile väljatrükitud transkriptsioonidena ja lintidena. Seda korpust on uurimiseks kasutatud mitmes valikus. Varaseim uuringute jaoks kasutatud korpus oli kogu aastal 1999 olemas olnud materjal, millest on tehtud esimesed statistilised analüüsid (Korpus 1999). Selle tulemusi on kasutatud Tiit Hennoste loengutes ning mitmetes artiklites (Hennoste 1999, 2000c).

Praegu on materjalid kasutatavad lisaks tekstidele kahe sõnaloendina. Esimeseks on Statistikakorpus 2000. Selle jaoks on valitud 2000. aasta kevadel olemas olevast korpusest 103 000 sõnaline alamkorpus, kus on püütud tasandada kopuse hetkeseisust johtuvat kallutatust. Kesksed valikukriteeriumid on järgmised:

- kui kogukorpuses on hulk samasuguseid tekste samalt inimeselt (nt sama inimene eri kauplustes), siis on osa neist välja jäetud;
- kui korpuses on palju samatüübilisi tekste (nt müügivestlused), mis tegelikult on ühiskonnas üsna haruldased, siis on neist valitud ainult mõned;
- kuna korpuses on liiga palju Tartu naisüliõpilaste omavahelisi vestlusi, siis on osa neist välja jäetud.

Sellest korpusest on Internetis kättesaadav sagedusloend 1100 sagedasema sõnavormiga.

Teine valikkorpus on silmast-silma suhtlust sisaldav argisituatsioonide ja avalike situatsioonide tuumrühmade paralleelkorpus (Paralleelkorpus 2000), milles on umbes 52 000 tekstisõnaline argikorpus ja 38 000 sõnaline avalike tekstide korpus. Sellest on paigutatud koduleheküljele kummagi korpuse 300 sagedasimat sõna.

Osakorpuste tekstivalikud on teinud T. Hennoste, A. Rääbis ja L. Lindström. Kõik korpuste arvutianalüüsiks vajalikud programmid on koostanud L. Lindström.

Korpusetekstide kasutamiseks on välja töötatud tüüpleping, mille sõlmivad kasutaja ja korpuse administraator Andriela Rääbis. Korpusest on pandud internetti näiteid erinevat tüüpi tekstidest koos tausta kirjeldavate päistega, transkriptsioonireeglid ning taustakirjelduse reeglid. Korpus ei ole tervikuna internetti väljapandav eetilistel põhjustel.

3. Allkeelte võrdlus

Üks oluline tekstikorpuste uurimise meetod on sõnasageduste võrdlemine korpusest tehtud sagedusloendite abil ehk leksikostatistika (vt sagedussõnastike klassikalisi uurimistüüpe Tuldava 1977). Üks tavalisemaid on eriti sagedaste sõnade või sõnavormide võrdlemine eri allkeeltes. See lubab kõige paremini välja tuua erinevate allkeelte statistiliselt kesksed erinevused, mis määravad kõige tugevamalt ka selle allkeele erijooned keelekasutaja teadvuses. Sellist uurimist võib teha rangete matemaatiliste meetoditega (nt kasutades hii-ruut testi või faktoranalüüsi; vt ülevaadet meetodite kohta McEnery, Wilson 1997: 61–86) või kasutada statistilisi erinevusi mittematemaatiliseks analüüsiks. Me oleme valinud siinkohal viimase tee, kasutades korpustest tehtud tekstisõnade sagedusridasid materjalina, et leida olulisi erinevusi allkeelte vahel toetudes eri keeltes tehtud lingvistilistele uuringutele.

Käesolevas uurimuses võrdleme nelja paralleelkorpust: suulise argikõne, suulise avaliku kõne, 1990. aastate ilukirjanduse ja 1990. aastate ajakirjanduse korpuseid. Ajakirjanduskorpus sisaldab umbes 232 000 tekstisõna, ilukirjanduse korpus 366 500 tekstisõna, suulise argikõne korpuses on 52 000 tekstisõna ja avaliku kõne korpuses 38 000. Nende allkorpuste põhjal on tehtud sõnavormide sagedusloendid, mille esimesed sada sõnavormi on esitatud Lisas olevas tabelis. Kuna korpused on erineva suurusega, siis on nende võrdlemise lihtsustamiseks arvutatud iga sõnavormi esinemissagedus kogu korpusest. See on välja toodud promillides. Kõik sagedusloendid on puhastamata, st sõnavormide homonüümid on koos samas sõnavormis (*tee* = käskiv kõneviis sõnast *tegema*, jook, sõidukoht).

Järgnevas võrdleme nende nelja sagedussõnastiku vormivalikut. Meid huvitab see, millised on nende sõnavormide kasutamise selged statistilised erinevused ja mida saab nende kaudu vastava allkeele kohta järeldada.

Mida need allkeeled esindavad? Suuline argivestlus esindab prototüüpset suulist suhtlust ja igasuguse kultuuri suhtluse tuumosa. See on suhtlus, mida me kõik kasutame kõige enam ja omandame kõige varem ning mida me kõik valdame nii aktiivselt kui passiivselt. Selle suhtluse infoväärtus on tüüpiliselt madal, tema sisu on üsna üldistatud, afektiivne ja interaktsionaalne. Ta on spontaanne, sundimatu dialoog, milles osaleja põhieesmärk on olla kaasa haaratud, osaleda. Argisuhtlus koosneb tüüpiliselt lühikestest kõnevoorudest, mille vahel on pikemaid jutustavaid lõike, kus üks osaleja räägib mingit lugu. Tema olulisteks positiivseteks leksikaalseteks erijoonteks on näiteks inglise keeles Douglas Biberi faktoranalüüsi abil tehtud uurimuste põhjal taju- ja tunnetusverbid (*arvama* jms), verbi olevikuvormid, *tegema* ja *olema* verbid, esimese ja teise isiku personaalpronoomenid, näitavad ja umbmäärased asesõnad, mitmesugused partiklid (rõhupartiklid, pehmendajad, dialoogipartiklid, toimetamispartiklid), modaalsõnad, lühenenud sõnavormid (vt Biber 1988).

Suuline avalik suhtlus on seesmiselt küllalt heterogeenne, st ta ei moodusta ühtset allkeelt ja tema sisemisi erinevusi pole meie teada statistiliselt uuritud. Kui mõõta kirjalike tekstidega, on ta väga lähedane suulise argisuhtluse keelekasutusele, st tema kesksed erijooned on samad kui argivestluses. Samas on selge, et ta on kirjalikule suhtlusele veidi lähem kui argisuhtlus.

Kirjalik ajakirjandus on valdkond, mille keel jaguneb Biberi järgi mitmesse registrisse, kuid enamjaolt kuulub sellisesse kirjalikku allkeelde, mida võib pidada statistiliselt neutraalseks, st temas ei ole grammatilisi ja leksikaalseid jooni, mille sagedus oleks allkeelte keskmisest kasutussagedusest oluliselt suurem või väiksem. Selline allkeel haarab lisaks ajakirjandusele ka muid valdkondi (populaarteadus, elulood, hobi- ja harrastusraamatute tekstid jms). Seega vastab see sõnarühm kõige enam sellele kirjakeele allkeelele, mida näiteks soome stiiliuurijad on nimetanud normaalproosaks ning milles on nähtud kirjakeele tuuma. Seega lubab argikeele ja ajakirjanduskeele võrdlemine meil laias laastus võrrelda suulise ja

kirjaliku keelekasutuse tuumosasid, kuigi selline määratlus pole rangelt võttes päris korrektne.

Ilukirjandus ehk kujutluslik narratiiv esindab tüüpiliselt omaette allkeelt, mis on üldiselt suulisele suhtlusele kõige lähem kirjalik allkeel, kuna kirjanduses kasutatakse ka dialoogi, mis on seotud suhtlejate ümber oleva situatsiooniga, ta on sama konkreetne kui suuline kõne ja temas on enam keelelist vabadust kui suhteliselt kitsaid norme järgivas neutraalstiilis. Ta erineb suulisest kõnest selle poolest, et tegu pole suulise ja spontaanse vaid kirjaliku ja redigeeritud tekstiga. Tema kõige olulisemaks statistiliseks keeleliseks erijooneks on Biberi järgi verbi minevikuvormide ja kolmanda isiku pronoomenite suur osakaal. Seega lubab avaliku suulise suhtluse ja ilukirjanduskeele võrdlemine meil võrrelda laias laastus neid suulise ja kirjaliku suhtluse allkeeli, mis võiksid olla ootuste järgi teineteisele kõige lähemal.

Alguses võrdleme omavahel kahte allkeeltepaari: suulist argisuhtlust ja kirjalikku ajalehekeelt ning suulist ametisuhtlust ja kirjalikku ilukirjanduskeelt. Seejärel vaatleme kokkuvõtlikult suulise ja kirjaliku keelevormi erinevusi. Lõpuks võrdleme omavahel suulist argisuhtlust ja suulist ametlikku suhtlust. Võrdluse aluseks on see, kui palju on ühes või teises allkeeles sõnu või sõnarühmi, mille kasutuses on selgeid statistilisi erinevusi.

3.1. Argikõne ja ajalehekeel

Sõnad, mis esinevad argikeeles oluliselt suurema sagedusega, on järgmised:

- dialoogipartiklid: *aa, ah, ahah, jaa, jah, mh, mhmh, mm*;
- piiripartiklid: *kule, no, sis, vä, onju*;
- toimetamispartiklid: *e, ee, mm, noh*;
- sidesõnad ja partiklid: *et, või*;
- adverbid ja partiklid: *nagu, nii*;
- rõhupartiklid: *ju, küll, ära, üldse, muidugi, ikka*;
- numbrid: *kaks, üks*;
- üldised koha ja aja proadverbid: *seal, sinna, siin, siis*;
- naer: *hehe*;
- isikulised asesõnad: *mina, minu, mul, sa, sul, tal, ma, me*;
- muud asesõnad: *see, mis, midagi, seda, mingi*;
- *olema: oled, olen, oli, olid, on*;
- eitus: *ei (ei)tea, (ei) ole*;

- lühenenud vormid: *aa 'ahah', i 'ei', kule, nimodi, sis, s 'siis' vä, a 'aga', mh 'mhmh*

Ajakirjanduskeeles on sagedasemad:
- nimisõnad: *aasta, eesti, krooni, liidu, Pärnu, Tallinna, raha, ajal*;
- adverbid: *ainult, enam, just, näiteks, siiski, vaid*;
- konkreetsemad koha ja ajasõnad: *ette, ajal, kohta, pärast, praegu, tagasi, üle, vastu*;
- isikulised asesõnad: *meie, tema, nende*;
- muud asesõnad: *iga, kes, oma, mille, sellest*;
- sidesõnad: *kuid, sest, ning*;
- verbivormid: *olla, olnud, peab, pole, ütles, võib.*

Mida saab sellest järeldada?

Argikorpuses esinevad suurema sagedusega sõnad, mis iseloomustavad suulist spontaanset dialoogi üldse: partiklid, üneemid, lühenenud vormid, üldised koha- ja ajaadverbid, eitus ja naer. Ajakirjanduskeeles on sagedasemad nimisõnad. Selgelt on näha ka see, millest ajakirjandus 1990. aastatel palju kõneles: rahast ja sellega seonduvast. Samuti on ajakirjanduskeeles sagedasemad konkreetsemad koha- ja ajasõnad (need sõnad võivad olla ka ühendverbide osad või adpositsioonid, vt nende kohta ka Hennoste, Muischnek 2000).

Veel mõned sõnad lubavad konkreetset tõlgendamist: *sest* ja *kuid* esinevad põhjuslausetes ning vastandust esitavates lausetes, *mille* ja *kes* on osalauseid siduvad sõnad. Mõlemad osutavad samuti põimlausete suuremale osakaalule ajalehekeeles. *ütles* on põhiline otsese kõne saatelause verb ajalehes. *ning* ja *pole* on *ja* ja *ei ole* sünonüümid, mida suuline kõne tarvitab väga vähe ja mida avalikus redigeeritud kirjalikus tekstis kasutatakse stilistilistel eesmärkidel. Kirjatekstide taotletakse teadlikult vaheldust, sünonüümide kasutamist, samal ajal kui suuline kõne eelistab samade sõnade kordust.

3.2. Avalik kõne ja ilukirjanduskeel

Avaliku suulise kõne ja ilukirjanduskeele põhierinevused on samasugused kui argikeele ja ajakirjanduskeele omad. Avalik suuline kõne kasutab sagedamini sõnarühmi, mis seostuvad otseselt suulise kõne erijoontega: toimetamispartiklid (*e, ee, hh, noh, õ, õõ*), mõned piiripartiklid (*eksole, no, vä*), lühemad mittekirjakeelsed sõnavormid

(*ned, nüd, se, sis, vä*), üldised koha- ja ajaproadverbid (*siin, seal, siis, nüüd*), numbrid, eitus.

Teiseks, sagedasemad on ka dialoogipartiklid ja muud suhtlussõnad (*ahah, mhmh, jah, jaa, aitäh, mm, palun, tere*), kuigi nende suuremat kasutust võiks eeldada ka ilukirjanduse dialoogis. Siiski näitab korpus, et ilukirjanduse dialoog on koostatud põhimõtteliselt teisiti kui suulise kõne dialoog. Samuti on suulises kõnes enam modaale: *saab, vaja, tuleb.*

Ilukirjanduses kasutatakse enam mitmesuguseid konkreetsemaid koha- ja ajasõnu (*ees, juurde, läbi, poole, tagasi, alla, pärast, vastu, üle, välja*) ja nimisõnu (*ema, elu, mees, naine*). Kuid on näha, et ilukirjanduse sagedased nimisõnad erinevad põhimõtteliselt ajakirjanduse kesksetest nimisõnadest. Ka siin on kirjatekstis enam kasutusel sõnad *pole/polnud* ja *ning.*

Huvitav on asesõnade jaotus. Suuline ametlik kõne kasutab enam *mina*, *see* ja *need* sõnu, ilukirjandus aga *meie, tema, nad* ja *sa* sarju. Seda on raske seletada. Eripärane on ka *olema* verbi vormide jaotus, kus olulisim on see, et kõne kasutab enam olevikku ja kiri minevikku. See on hästi kooskõlas suulise suhtluse ja kirjaliku suhtluse üldiste ajakasutuse tendentsidega.

3.3. Kõne ja kiri

Eelnevast võib näha, et põhierinevused on ka eesti keeles ikka suulise (spontaanse) ja kirjaliku (redigeeritud) kõne sõnakasutuse vahel. Seda ei muuda ka ilukirjanduse rohke dialoog ega situatsiooni argisus või avalikkus. Kui teeme kokkuvõtte suulise ja kirjaliku teksti kesksetest erinevustest nende allkeelte põhjal, siis saame välja tuua mitmed olulised statistilised erinevused. Esiteks, suuline kõne kasutab sagedamini mitmeid sõnarühmi:

- lühemad ja selliselt mittekirjakeelsed vormid (*ned, nüd, se, sis, vä*);
- üneemid ja toimetamispartiklid (*e, ee, hh, noh, õ, õõ*);
- suhtluspartiklid ja suhtlussõnad (*ahah, mhmh, jah, jaa, aitäh, mm, palun, tere*), mis eristavad antud juhul dialoogilist suulist kõnet valdavalt monoloogilisest kirjatekstist. Nagu öeldud, ei muuda seda ka ilukirjanduse dialoog;
- suulises suhtluses on enam mitmesuguseid muid partikleid: (*eksole, no, vä*);

- suulises suhtluses on enam numbreid, mille sagedasem kasutus on osalt seotud avalikus korpuses olevate situatsioonide teemadega (kauplused jms), osalt raskesti tõlgendatav;
- suulises kõnes kasutatakse enam hästi üldisi koha ja aja proadverbe (*siin, seal, siis, nüüd*);
- suuline kõne kasutab enam *ei* sõna, mis seostub suulises kõnes eelkõige eitavas kõnelaadis verbidega, mitte eitavate vastustega;
- suulises kõnes enam modaale: *saab, vaja, tuleb.*

Teiseks, kirjas on enam kolme liiki sõnu:
- kirjas kasutatakse enam mitmesuguseid konkreetsemaid koha- ja ajasõnu (*ees, läbi, tagasi, pärast, vastu, üle, välja*), mis võivad olla ka positsioonid või ühendverbi osad;
- kasutatakse enam nimisõnu, mis aga erinevad ilukirjandus- ja ajakirjanduskeeles;
- samuti on sõnad *pole/polnud* ja *ning* tugevalt kirjaliku keele sõnad. Selle taga on kirjaliku suhtluse stiilitaotelused, mis käsivad vaheldada sagedaste sõnade sünonüüme.

Kolmandaks, kirjalike tekstide esimesed sada sõna katavad 32–34% tekstidest, suulise korpuse sada sõna aga 46–47%. Seega on suulise kõne sõnade tekstikatmus oluliselt suurem, ehk suulises kõnes saadakse hakkama väiksema arvu sõnadega ja kasutatakse vähem sünonüüme.

Neljandaks, mõned asesõnad ja verb *olema* jaotuvad eri allkeelte vahel eripäraselt:
- suuline ametlik kõne kasutab enam *mina, see* ja *need* sõnu, argikõne *mina* ja *sina*, ilukirjandus kasutab enam *meie, tema* sarju ja ajakirjandus *meie*, *tema* ja *nende* sarju. Suuline kõne on selgelt minakesksem. Seda on näidanud ka teiste keelte uurimused. Teisalt, kirjalikud tekstid on siin suhteliselt ühesugused, kuid suulised erinevad omavahel. Vahe on siin ka argise ja avaliku suhtluse vahel. Avalik suhtlus kasutab enam kolmanda isiku sõnu, st räägib enam kellestki või millestki;
- *olema* verbi vormide sagedused olenevad konkreetsest allkeelest. Üldine erinevus on, et kõne kasutab enam eitusele või käsule viitavat vormi *ole*.

3.4. Ametlik ja argine suuline kõne

Mille poolest erinevad omavahel argine ja avalik suuline kõne? Sõnavormid, mida on argisuhtluses selgelt enam kui avalikus suulises suhtluses, on järgmised:

- mitmesugused lühemad sõnavormid: *aa 'ahah', i 'ei', kule, ku 'kui', nimodi, sis, s 'siis', vä, a 'aga', mh 'mhmh'*;
- erinevad interjektsioonilised partiklid, mis väljendavad mitmesuguseid emotsioone: *ah, hehe* (naer);
- takerduspartikkel: *noh*;
- piiripartiklid, mis alustavad ja lõpetavad kõnevoore ning lausungeid ja osutavad eelkõige üksustevaheliste seoste tüüpi: *ja, sis/s, no, vä, a/aga, kule, onju, tead, vaata*;
- rõhutavad partiklid: *ära, muidugi, ju, ikka, küll, üldse*;
- umbmäärastavad või pehmendavad partiklid: *seal, vist, mingi*;
- isikulised asesõnad, eriti *mina* ja *sina*: *minu, ma, sul, tal, tema, ta, sa*;
- *olema* verbi mõned vormid: *oled, olen, olid, oli;*
- eitus: *ei, (ei) saa, (ei) tea*, mis seostub eelkõige eitava kõnelaadiga mingis pikemas tekstis, mitte eitavate vastustega.

Avalikus suhtluses kasutatakse enam järgmisi sõnarühmi:

- suhtlusrutiinisõnad (tänamised, palumised jms): *aitäh, palun, tere*;
- takerdusüneemid: *hh, õ/õõ, ee/e;*
- dialoogipartiklid: *jah, mhmh, ahah, selge;*
- *olema: olema, on;*
- numbrid ja muid hulki ja väärtusi osutavad sõnad: *üks, palju, väga, kaks, kolm, viis, krooni;*
- küsivad sõnad: *kas, eksole, kui;*
- mõned isikulised asesõnad: *te, teil, teile, me, meil, meie;*
- osutavad asesõnad, aseadverbid ja üldsõnad: *asi, ned, nüd, siit, need, nüüd, siin.*

Kuidas neid erinevusi tõlgendada?

- Argivestlus on temaatiliselt rohkem inimkeskne (*tema*) ja ametivestlus asjakeskne (osutavad asesõnad), aga ka numbrid ja muud hulki osutavad sõnad. Kindlasti on see seotud ka kauplusedialoogide rohkusega avalikus suhtluses.
- Argisuhtlus on tugevalt *mina-sina* keskne suhtlus, kus püütakse aktiivselt tähelepanu ning osutatakse enam endale ja partnerile kui

ametivestluses. Seda osutab nii asesõnade kasutus kui ka mitmesugused partiklid, mis püüavad tähelepanu ja nõusolemist (*kule, onju, vä, tead, vaata*).

- Argisuhtlus on emotsionaalsem, seda osutab emotsionaalsete sõnade suurem osakaal aga ka rõhutavate partiklite suurem osakaal.
- Argisuhtlus on spontaansem, mida osutab see, et ta kasutab enam pehmendamist ja umbmäärastamist.
- Argisuhtlus koosneb pikematest kõnevoorudest või teemaarendustest, milles on enam omavahel seotud lausungeid ja mõttepöördeid. Seda osutavad piiripartiklid *no, ja, sis*.
- Argisuhtlus kasutab enam sõnade lühivorme, mis pole kirjakeelsed ning mida seetõttu avalikus suhtluses välditakse.
- Avalik suhtlus kasutab enam asesõnu *teie* ja *meie*. Mõlemad neist aitavad hoida ja luua koostööd. Samas suunas viitab see, et avalik vestlus kasutab vähem eitamist, mis on üldjuhul koostööd rikkuv strateegia, normivastane suhtlusvõte, mida üritatakse vältida. Samale viitavad ka mitmesugused tänamised jms viisakussõnad. Viimaste suur osakaal on seotud ka avaliku suhtluse lühemate tekstidega, kuhu sisse läinud rohkem suhtluse algusi ja lõppe.
- Avalik suhtlus eelistab takerdumisel kasutada üneeme ja argisuhtlus näiteks sagedasimat takerdupartiklit *noh*. Selline takerdusvahendite erinev eelistamine on tuntud ka muudest keeltest ja selle taustal on asjaolu, et üneemid torkavad vähem kõrva ning loovad mulje sujuvamast jutust.
- Avalik suhtlus kasutab dialoogipartikleid enam kui argisuhtlus. Lisaks kasutab avalik suhtlus partiklite täisvorme enam kui argisuhtlus ning partiklite lühemaid vorme vähem. Nii moodustab *ahah* avalikus korpuses 3,9‰ sõnavormidest, argikorpuses 1,3‰. Lühenenud variant *aa* aga annab avalikus korpuses 0,9‰ ja argikorpuses 2,1‰. *mhmh* moodustab avalikus korpuses 9,7‰, argikorpuses 6,2‰. Lühike vorm *mh* annab avalikus korpuses 0,5‰, argikorpuses 1,1‰. *jah* moodustab avalikus korpuses 15‰, argikorpuses 12,4‰. Selle taustamõjuriks võivad olla nii kühemad kõnevoorud avalikus suhtluses kui ka rohkem küsimusi, seisukohavõtte, palveid jms, ühesõnaga suhtluse suurem infokeskus.
- Avalikus vestluses on enam osutavaid sõnu, mis viitavad situatsioonile ning objektidele, millest kõneldakse (*asi, need,*

siin, siis). Argisuhtluses on vastupidi enam isikutele viitavaid asesõnu, eriti *mina* ja *sina* sõna. Seegi näitab avaliku suhtluse suuremat infokesksust.

Kirjandus

Biber, D. 1988. Variation across Speech and Writing. Cambridge: Cambridge UP

Chafe, W., Tannen, D. 1987. The relation between written and spoken language. – Annual Review of Anthropology 16, 383–407.

Chafe, W 1982. Integration and involvement in speaking, writing and oral literature. – Spoken and Written Language. Exploring Orality and Literacy. Toim D. Tannen. Norwood, N.J.: Ablex. 35–53.

Crowdy, S. 1993. Spoken corpus design. – Literary and Linguistic Computing Vol 8 (4), 259–264.

Du Bois, J. W. 1991. Transcription design principles for spoken discourse research. – Pragmatics: Quarterly Publication of the International Pragmatics Association 1(1), 71–106.

Du Bois, J. W., Scuetze-Coburn, S., Cumming, S., Paolino, D. 1993. Outline of discourse transcription. – Talking Data: Transcription and Coding in Discourse Research. Toim J. A. Edwards, M. D. Lampert. Hillsdale, NJ: Lawrence Erlbaum. 45–89.

Edwards, J. 1992. Transcription of discourse. – International Encyclopedia of Linguistics. Toim W. Bright. Vol 1. New York, Oxford: Oxford UP. 367–371.

Edwards, J. 1993a. Principles and contrasting systems of discourse transcription. – Talking Data: Transcription and Coding in Discourse Research. Toim J. A. Edwards, M. D. Lampert. Hillsdale, NJ: Lawrence Erlbaum. 3–31.

Edwards, J. 1993b. Survey of electronic corpora and related resources for language researchers. – Talking Data: Transcription and Coding in Discourse Research. Toim J. A. Edwards, M. D. Lampert. Hillsdale, NJ: Lawrence Erlbaum. 267–310.

Edwards, J. 1995. Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. – Spoken English on Computer. Transcription, Mark-up and Application. Toim G. Leech, G. Myers, J. Thomas. London: Longman. 19–34.

Hennoste, T. 1996. Tartu University Corpus of Written Estonian: A survey of the structure of texts and principles of selection. – Estonian in the Changing World. Toim H. Õim. University of Tartu. Dept. of General Linguistics. 7–32.

Hennoste, T. 1999. Suulise kõne sõnakasutuse sagedased erijooned. – 75 vuotta viroa Helsingin yliopistossa. Viron kielen ja kultuurin opettaminen Suomessa-seminaari 23. 11. 1998. Castrenianumin toimitteita 56. Toim R. Kasik ja L. Huima. Helsinki. 98–131.

Hennoste, T. 2000a. Eesti suulise kõne uurimine: transkriptsioon, taust ja korpus. – Keel ja Kirjandus 2, 91–106.

Hennoste, T. 2000b. Sissejuhatus suulisesse eesti keelde I. Taust ja uurimisobjekt. – Akadeemia 5, 1117–1150.

Hennoste, T. 2000c Sissejuhatus suulisesse eesti keelde II. Suulise kõne erisõnavara I. – Akadeemia 6, 1343–1374.

Hennoste, T. 2000d Sissejuhatus suulisesse eesti keelde III. Suulise kõne erisõnavara II. – Akadeemia 7, 1553-1582.

Hennoste, T. 2000e Sissejuhatus suulisesse eesti keelde IV Suulise kõne erisõnavara III. Partiklid. – Akadeemia 8, ilmumas.

Hennoste, T., Koit, M., Roosmaa, T., Saluveer, M. 1998. Structure and usage of the Tartu University Corpus of Written Estonian. – International Journal of Corpus Linguistics. Vol 3 (2), 279–304.

Hennoste, T., Muischnek, K. 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Käesolevas kogumikus.

Hutchby, I., Wooffitt, R. 1998. Conversation Analysis. Polity Press.

Johansson, S. 1995. The approach of the Text Encoding Initiative to the encoding of spoken discourse. – Spoken English on Computer. Transcription, Mark-up and Application. Toim G. Leech, G. Myers, J. Thomas. London: Longman. 82–98.

Leech, G., Myers, G., Thomas, J. (toim) 1995. Spoken English on Computer. Transcription, Mark-up and Application. London: Longman.

McEnery, T., Wilson, A. 1997. Corpus Linguistics. Edinburgh: Edinburgh UP

Muischnek, K. 1998. Korpuslingvistika. – Keel ja Kirjandus 1, 8–12.

Ochs, E. 1979, Transcription as theory. – Developmental Pragmatics. Toim E. Ochs and B. Schiefflen. N.Y.: Academic Press. 43–72.

Pajusalu, K. 1997. Keskse perifeeria mõjust eesti keele tekkeloos. – Pühendusteos Huno Rätsepale. Tartu Ülikooli eesti keele õppetooli toimetised 7. Toim M. Erelt, M. Sedrik, E. Uuspõld. Tartu. 167–184.

Peppé, S. 1995. The survey of English usage and the London–Lund Corpus: computerising manual prosodic transcription. – Spoken English on Computer. Transcription, Mark-up and Application. Toim G. Leech, G. Myers, J. Thomas. London: Longman. 187–202.

Russkaja razgovornaja retsh. Tekstõ. Moskva, 1978.

SICLRE II 2000 = Second International Conference on Language Resources and Evaluation. Athens, Greece 31. May – 2. June. Proceedings, Vol 2. 877-982.

Sinclair, J. 1995. From theory to practice. – Spoken English on Computer. Transcription, Mark-up and Application. Toim G. Leech, G. Myers, J. Thomas. London: Longman. 99–109.

Suojanen, M. K. 1985. Mitä Turussa puhutaan. Raportti Turun puhekielen tutkimuksesta. Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja 23, Turku: Turun Yliopisto.

Svartvik J. (toim) 1990. The London–Lund Corpus of Spoken English. Description and Research. Lund: Lund UP.

Svartvik, J., Quirk, R. (toim) 1980. A Corpus of English Conversation. Lund Studies in English 56. Lund: Lund UP.

Zemskaja, E. A. 1979. Russkaja razgovornaja retsh. Moskva: Ruskij Jazõk.

Tainio, L. (toim) 1997. Keskustelunanalyysin perusteet. Tampere: Vastapaino.

Tuldava, J. 1977 Sagedussõnastik leksikostatistilise uurimise objektina. – Keelestatistika 2. TRÜ Toimetised 413. Tartu: TRÜ. 141–169.

Lisa. 100 sagedasimat sõnavormi neljas allkeeles

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
on	9612	25,43	ei	10372	16,93	on	1634	31,36	on	1644	43,06
et	4368	11,56	ta	9972	16,28	ja	1569	30,11	ei	916	23,99
ei	3864	10,22	on	8380	13,68	ei	1477	28,34	ja	896	23,47
ka	3096	8,19	oli	8218	13,42	et	1176	22,57	et	868	22,74
kui	2843	7,52	et	7541	12,31	ma	913	17,52	see	679	17,78
eesti	2351	6,22	kui	5835	9,53	see	850	16,31	jah	572	14,98
oli	1963	5,19	ma	4911	8,02	jah	648	12,44	ma	494	12,94
oma	1899	5,02	see	4852	7,92	oli	618	11,86	ka	391	10,24
see	1858	4,92	aga	4352	7,10	ta	593	11,38	mhmh	369	9,66
ning	1739	4,60	oma	3827	6,25	noh	545	10,46	siis	336	8,80
aga	1728	4,57	siis	3656	5,97	aga	509	9,77	kui	333	8,72
mis	1522	4,03	mis	3245	5,30	sis	499	9,58	ta	314	8,22
ta	1369	3,62	ka	2993	4,89	siis	430	8,25	nii	310	8,12
nii	1243	3,29	nagu	2815	4,60	ka	427	8,19	aga	310	8,12
või	1206	3,19	nii	2604	4,25	mis	419	8,04	noh	302	7,91
kes	1174	3,11	ning	2416	3,94	seal	399	7,66	ee	286	7,49
siis	1142	3,02	seda	2397	3,91	sa	365	7,00	või	281	7,36
seda	1069	2,83	või	2255	3,68	no	365	7,00	mis	274	7,18
pole	1036	2,74	tema	2238	3,65	nii	345	6,62	ole	259	6,78
selle	983	2,60	veel	2025	3,31	kui	327	6,28	oli	242	6,34
meie	975	2,58	sa	1967	3,21	mhmh	324	6,22	siin	218	5,71
veel	804	2,13	kes	1938	3,16	või	310	5,95	no	216	5,66
tema	758	2,01	nad	1851	3,02	seda	304	5,83	sis	214	5,61

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
võib	723	1,91	kas	1839	3,00	hehe	303	5,81	kas	212	5,55
juba	721	1,91	midagi	1807	2,95	ära	287	5,51	e	207	5,42
kuid	700	1,85	kõik	1624	2,65	ole	277	5,32	need	194	5,08
välja	687	1,82	nüüd	1612	2,63	ju	268	5,14	se	186	4,87
kas	686	1,81	ära	1611	2,63	vä	255	4,89	te	171	4,48
tuleb	668	1,77	juba	1606	2,62	ee	244	4,68	seda	169	4,43
kus	658	1,74	selle	1558	2,54	ikka	236	4,53	seal	166	4,35
vaid	657	1,74	olid	1481	2,42	mina	235	4,51	nüüd	162	4,24
nende	635	1,68	välja	1425	2,33	nagu	231	4,43	selle	158	4,14
ole	631	1,67	pole	1422	2,32	se	215	4,13	nagu	156	4,09
nagu	628	1,66	mitte	1290	2,11	a	197	3,78	ära	154	4,03
üle	625	1,65	küll	1289	2,10	jaa	194	3,72	ahah	148	3,88
kõik	621	1,64	mida	1278	2,09	selle	193	3,70	me	134	3,51
mitte	598	1,58	kuid	1269	2,07	mingi	185	3,55	üks	128	3,35
a	586	1,55	teda	1266	2,07	küll	178	3,42	kõik	127	3,33
mida	582	1,54	polnud	1243	2,03	tea	171	3,28	jaa	127	3,33
eest	576	1,52	oleks	1243	2,03	onju	170	3,26	midagi	125	3,27
aasta	566	1,50	enam	1156	1,89	midagi	170	3,26	väga	119	3,12
vastu	554	1,47	üle	1154	1,88	kõik	170	3,26	nad	104	2,72
ma	528	1,40	ütles	1149	1,88	kas	170	3,26	juba	97	2,54
üks	516	1,37	ju	1145	1,87	need	163	3,13	veel	96	2,51
oleks	508	1,34	sest	1121	1,83	üks	153	2,94	hh	94	2,46
olnud	500	1,32	siin	1104	1,80	nüüd	143	2,74	teil	93	2,44

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**
nad	500	1,32	ainult	1082	1,77	me	140	2,69	ja:	92	2,41
väga	497	1,31	olla	1061	1,73	veel	134	2,57	kaks	91	2,38
nüüd	494	1,31	olnud	1056	1,72	tuleb	134	2,57	tuleb	89	2,33
palju	492	1,30	ega	1048	1,71	siin	132	2,53	meil	89	2,33
pärast	480	1,27	mees	1041	1,70	mul	132	2,53	neid	87	2,28
need	480	1,27	vastu	1027	1,68	nad	123	2,36	viis	86	2,25
ainult	477	1,26	nende	1018	1,66	kes	123	2,36	ikka	86	2,25
sest	471	1,25	peale	1010	1,65	juba	123	2,36	küll	85	2,23
saab	462	1,22	minu	1009	1,65	vist	122	2,34	mul	84	2,20
praegu	460	1,22	kuidas	999	1,63	e	116	2,23	ju	84	2,20
olid	459	1,21	ole	993	1,62	olid	115	2,21	palju	83	2,17
kõige	448	1,19	mu	988	1,61	tead	111	2,13	õ	82	2,15
kohta	440	1,16	seal	961	1,57	aa	110	2,11	tea	82	2,15
ära	440	1,16	vaid	960	1,57	neid	108	2,07	a	81	2,12
eestis	438	1,16	pärast	957	1,56	väga	102	1,96	peale	79	2,07
aastal	429	1,13	talle	952	1,55	sinna	97	1,86	ned	79	2,07
olla	422	1,12	tal	948	1,55	kus	93	1,78	aitäh	76	1,99
peab	417	1,10	tagasi	936	1,53	i	93	1,78	teha	75	1,96
enam	416	1,10	neid	934	1,52	palju	92	1,77	vist	74	1,94
ajal	414	1,10	meie	928	1,51	ega	90	1,73	palun	73	1,91
küll	413	1,09	mulle	923	1,51	peale	87	1,67	oma	71	1,86
me	412	1,09	ise	912	1,49	tal	85	1,63	õõ	70	1,83
liidu	404	1,07	ikka	912	1,49	sul	85	1,63	sa	70	1,83

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
tallinna	394	1,04	need	906	1,48	mitte	85	1,63	nüd	69	1,81
kuidas	394	1,04	üks	906	1,48	välja	83	1,59	mina	69	1,81
mille	379	1,00	kus	902	1,47	minu	83	1,59	tere	68	1,78
neid	378	1,00	mind	893	1,46	oma	80	1,54	saab	68	1,78
midagi	375	0,99	mina	885	1,44	meil	80	1,54	eksole	67	1,75
tagasi	374	0,99	me	869	1,42	hea	80	1,54	välja	66	1,73
kogu	370	0,98	mul	847	1,38	täna	79	1,52	kes	66	1,73
raha	366	0,97	tuli	832	1,36	saa	79	1,52	siit	63	1,65
ise	364	0,96	läbi	822	1,34	olen	79	1,52	meie	63	1,65
ju	356	0,94	olen	816	1,33	ah	79	1,52	vaja	61	1,60
osa	354	0,94	poole	815	1,33	mm	78	1,50	võib	59	1,55
sellest	351	0,93	sellest	743	1,21	ku	77	1,48	selline	59	1,55
siin	348	0,92	võib	739	1,21	kuidas	75	1,44	selge	59	1,55
rohkem	346	0,92	keegi	716	1,17	kohe	75	1,44	mitte	57	1,49
peale	342	0,90	palju	701	1,14	üldse	74	1,42	krooni	56	1,47
ega	338	0,89	ema	688	1,12	oleks	73	1,40	lihtsalt	55	1,44
teha	322	0,85	isegi	684	1,12	teha	71	1,36	ega	55	1,44
saa	320	0,85	naine	677	1,11	nimodi	71	1,36	sinna	54	1,41
siiski	316	0,84	väga	675	1,10	ise	71	1,36	sest	54	1,41
näiteks	316	0,84	end	675	1,10	s	70	1,34	oleks	54	1,41
meil	313	0,83	ette	672	1,10	tema	69	1,32	vä	53	1,39
kaks	308	0,81	alla	672	1,10	mida	69	1,32	olema	53	1,39
krooni	306	0,81	eest	669	1,09	kule	69	1,32	näiteks	53	1,39

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
ette	306	0,81	just	661	1,08	ahah	69	1,32	kolm	53	1,39
just	301	0,80	miks	643	1,05	saab	68	1,30	just	53	1,39
läbi	299	0,79	saanud	636	1,04	muidugi	67	1,29	mm	52	1,36
aastat	297	0,79	tuleb	624	1,02	kaks	64	1,23	mida	51	1,34
iga	292	0,77	ees	619	1,01	tuli	62	1,19	teile	50	1,31
ütles	290	0,77	juurde	608	0,99	mh	59	1,13	kus	50	1,31
pärnu	284	0,75	kõige	599	0,98	vaata	58	1,11	asi	50	1,31
seal	283	0,75	elu	595	0,97	oled	58	1,11	kuidas	49	1,28
i	283	0,75	jälle	586	0,96	ja:	57	1,09	olnud	48	1,26
kokku	281	0,74	all	582	0,95	sulle	56	1,07	ise	48	1,26
saanud	280	0,74	kohe	580	0,95	mhemhe	56	1,07	ütleme	47	1,23
poolt	280	0,74	hakkas	564	0,92	ütles	54	1,04	tähendab	47	1,23
nõukogude	280	0,74	enne	557	0,91	te	54	1,04	sellest	47	1,23
tartu	276	0,73	kogu	555	0,91	taha	54	1,04	olla	47	1,23
vene	275	0,73	te	544	0,89	siss	54	1,04	üle	46	1,20
sai	274	0,72	vahel	533	0,87	pole	54	1,04	siia	46	1,20
koos	274	0,72	edasi	526	0,86	jälle	54	1,04	sellepärast	46	1,20
vabariigi	272	0,72	üles	525	0,86	aru	54	1,04	peab	46	1,20
ühe	272	0,72	olin	521	0,85	sealt	53	1,02	on:	46	1,20
poole	270	0,71	sai	514	0,84	nojah	53	1,02	läheb	46	1,20
kuni	266	0,70	läks	513	0,84	mulle	53	1,02	selles	45	1,18
euroopa	266	0,70	kaks	511	0,83	sest	52	1,00	rohkem	45	1,18
isegi	265	0,70	jäi	511	0,83	sellest	52	1,00	olid	45	1,18

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
tuli	258	0,68	maha	510	0,83	miks	52	1,00	muidugi	45	1,18
peaks	257	0,68	lahti	510	0,83	meie	52	1,00	praegu	44	1,15
enne	251	0,66	jah	508	0,83	enam	52	1,00	nimodi	43	1,13
alla	250	0,66	sisse	505	0,82	aeg	52	1,00	hehe	43	1,13
n	249	0,66	saa	505	0,82	tähendab	50	0,96	eks	43	1,13
vaja	246	0,65	teha	504	0,82	pärast	50	0,96	ainult	43	1,13
saada	245	0,65	muidugi	503	0,82	olnud	49	0,94	nojah	42	1,10
maa	244	0,65	isa	501	0,82	asi	48	0,92	pärast	41	1,07
nsv	242	0,64	kord	494	0,81	õõ	47	0,90	ni	41	1,07
soome	239	0,63	tea	485	0,79	siuke	47	0,90	mingi	41	1,07
olema	239	0,63	oled	480	0,78	sina	47	0,90	enam	41	1,07
ehk	239	0,63	pidi	478	0,78	lihtsalt	47	0,90	neli	40	1,05
selles	238	0,63	mille	476	0,78	ütleb	46	0,88	kuus	40	1,05
eriti	235	0,62	aeg	463	0,76	teda	46	0,88	aa	40	1,05
rubla	234	0,62	alles	461	0,75	siia	46	0,88	sada	39	1,02
samuti	231	0,61	ennast	459	0,75	kah	46	0,88	natuke	39	1,02
balti	231	0,61	pea	455	0,74	võibolla	45	0,86	läbi	39	1,02
järgi	224	0,59	ajal	454	0,74	praegu	45	0,86	ühe	38	1,00
vahel	223	0,59	ümber	452	0,74	peal	45	0,86	vel	38	1,00
usa	222	0,59	taga	443	0,72	üts	44	0,84	kohe	38	1,00
muidugi	222	0,59	ometi	440	0,72	ä	44	0,84	isegi	38	1,00
minu	222	0,59	kinni	439	0,72	vaja	44	0,84	ikkagi	38	1,00
korda	222	0,59	aega	437	0,71	sääl	44	0,84	võibolla	37	0,97

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
ikka	222	0,59	hea	433	0,71	sisse	44	0,84	teie	37	0,97
suur	219	0,58	mööda	432	0,71	sellepärast	44	0,84	tal	37	0,97
olen	219	0,58	kunagi	432	0,71	päris	44	0,84	sin	37	0,97
juurde	218	0,58	aru	432	0,71	muidu	44	0,84	saa	37	0,97
alles	218	0,58	selles	431	0,70	minna	44	0,84	olen	37	0,97
võiks	214	0,57	enda	429	0,70	ilus	44	0,84	juurde	37	0,97
linna	214	0,57	koos	424	0,69	peab	43	0,83	öelda	35	0,92
edasi	212	0,56	võis	420	0,69	olla	43	0,83	peaks	35	0,92
riigi	211	0,56	siiski	420	0,69	kogu	43	0,83	olete	35	0,92
kuigi	209	0,55	vaatas	419	0,68	hästi	43	0,83	aga:	35	0,92
sel	208	0,55	ent	416	0,68	vot	42	0,81	päris	34	0,89
rootsi	208	0,55	kokku	414	0,68	viis	42	0,81	minu	34	0,89
lääne	207	0,55	vana	412	0,67	talle	42	0,81	kuna	34	0,89
elu	206	0,54	ehk	411	0,67	isegi	42	0,81	ette	34	0,89
s	204	0,54	rohkem	409	0,67	teed	41	0,79	üldse	33	0,86
kuna	204	0,54	praegu	407	0,66	tee	41	0,79	ää	33	0,86
end	201	0,53	peab	405	0,66	tagasi	41	0,79	i	33	0,86
töö	200	0,53	küsis	401	0,65	olnd	41	0,79	hästi	33	0,86
selleks	200	0,53	järele	399	0,65	olema	41	0,79	võtta	32	0,84
hea	200	0,53	taha	398	0,65	niimodi	41	0,79	tegelikult	32	0,84
jääb	199	0,53	teie	395	0,64	kolm	41	0,79	tagasi	32	0,84
venemaa	196	0,52	kuigi	393	0,64	rohkem	40	0,77	suur	32	0,84
teine	195	0,52	sina	389	0,64	nüd	40	0,77	mulle	32	0,84

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰	Sõna	Hulk	‰
sama	193	0,51	üldse	389	0,64	minema	40	0,77	kakskend	32	0,84
miks	191	0,51	päris	388	0,63	iga	40	0,77	sai	31	0,81
jooksul	191	0,51	suur	385	0,63	läheb	39	0,75	peal	31	0,81
võtta	190	0,50	endale	385	0,63	asja	39	0,75	neil	31	0,81
sõnul	190	0,50	alati	384	0,63	õ	38	0,73	enne	31	0,81
oleme	189	0,50	iga	383	0,63	keegi	38	0,73	teine	30	0,79
neist	189	0,50	ühe	382	0,62	juurde	38	0,73	saate	30	0,79
m	189	0,50	vist	376	0,61	ema	38	0,73	mõttes	30	0,79
võimalik	187	0,49	võttis	376	0,61	teine	37	0,71	mt	30	0,79
uus	187	0,49	olema	370	0,60	sai	37	0,71	eriti	30	0,79
valitsuse	186	0,49	teine	369	0,60	pidi	37	0,71	aeg	30	0,79
aega	186	0,49	saab	368	0,60	oi	37	0,71	vastu	29	0,76
all	185	0,49	sinna	367	0,60	nigu	37	0,71	töö	29	0,76
tuleks	183	0,48	korda	362	0,59	mingid	37	0,71	pool	29	0,76
tallinnas	182	0,48	tegi	356	0,58	täpselt	36	0,69	kindlasti	29	0,76
mina	179	0,47	minna	351	0,57	tegelikult	36	0,69	jälle	29	0,76
keegi	179	0,47	täis	349	0,57	mingit	36	0,69	tuhat	28	0,73
ilma	179	0,47	tulnud	347	0,57	aega	36	0,69	seitse	28	0,73
uue	176	0,47	juures	347	0,57	võib	35	0,67	m	28	0,73
aeg	176	0,47	su	345	0,56	tule	35	0,67	korda	28	0,73
kelle	175	0,46	mingi	345	0,56	saad	35	0,67	kohta	28	0,73
ees	175	0,46	äkki	339	0,55	ongi	35	0,67	tõesti	27	0,71
olevat	174	0,46	kuhu	337	0,55	natuke	35	0,67	tema	27	0,71

Ajakirjandus			Ilukirjandus			Suuline argisuhtlus			Suuline ametlik suhtlus		
Sõna	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**	**Sõna**	**Hulk**	**‰**
esimene	174	0,46	mõni	336	0,55	käis	35	0,67	sellega	27	0,71
tööd	172	0,46	sinu	335	0,55	kuskil	35	0,67	raha	27	0,71
sellele	172	0,46	sulle	334	0,55	hh	35	0,67	hea	27	0,71
kahe	170	0,45	jäänud	333	0,54	eks	35	0,67	vaatame	26	0,68
hästi	170	0,45	peaaegu	330	0,54	tõesti	34	0,65	siss	26	0,68
rahva	169	0,45	olevat	327	0,53	tean	34	0,65	nigu	26	0,68
puhul	169	0,45	silmad	326	0,53	taga	34	0,65	ilmselt	26	0,68
kolm	169	0,45	tüdruk	322	0,53	peaks	34	0,65	esimene	26	0,68
juures	169	0,45	inimene	315	0,51	läbi	34	0,65	võimalik	25	0,65
vähemalt	167	0,44	lihtsalt	313	0,51	kuradi	34	0,65	täna	25	0,65
kord	167	0,44	sind	312	0,51	inimene	34	0,65	sisse	25	0,65
ülem-nõukogu	165	0,44	lõpuks	312	0,51	aastat	34	0,65	sellist	25	0,65
seni	162	0,43	öelda	311	0,51	vastu	33	0,63	sealt	25	0,65
valitsus	161	0,43	tõesti	309	0,50	umbes	33	0,63	protsenti	25	0,65
kaasa	161	0,43	otsa	308	0,50	mees	33	0,63	muidu	25	0,65
maailma	160	0,42	hästi	306	0,50	väike	32	0,61	miks	25	0,65
ilmselt	160	0,42									

Konversatsiooniagendi modelleerimine*

Mare Koit, Haldur Õim
Tartu Ülikool

1. Sissejuhatus

Seoses märkimisväärsete edusammudega kõnetuvastuse ja -sünteesi alal (nt ingliskeelse kõne automaatse analüüsi ja sünteesi mooduleid müüakse juba koos kontoritarkvaraga) on viimastel aastatel oluliselt tõusnud huvi ka dialoogi modelleerimise vastu. 1998 loodi Arvutuslingvistika Assotsiatsiooni (ACL) spetsiaalne huvigrupp SIGdial, mis koondab diskursuse ja dialoogi töötlemisega tegelevaid isikuid ja vahendab vastavaid keeleressursse (http://www.sigdial.org/). Korraldatakse mitmeid rahvusvahelisi seminare, kus käsitletakse suhtlust arvutiga loomulikus keeles (sh Teise rahvusvahelise keeleressursside konverentsi LREC 2000 satelliitseminar).

Loomuliku keele automaattöötluse algaastatel uuriti dialoogi põhiliselt kahe ülesande raames – masintõlge ja küsimus-vastussüsteemid suhtlemiseks andmebaasidega. Kuid katsed luua loomuliku keele sisendit ja väljundit andmebaasisüsteemidele, selleks et analüüsida kasutaja küsimusi ja produtseerida kooperatiivseid vastuseid, jäid paljuski ainult uurimiseesmärgiks ega realiseerunud soovitud määral.

Praegu on taas tõusnud huvi kooperatiivsete dialoogide modelleerimise vastu, mille praktiliseks rakenduseks võivad olla näiteks interaktiivsed telefoni- või veebiteenused (sh süsteemid telefonikõnede ümbersuunamiseks, päästeteenistuse planeerimiseks, reisiinfo hankimiseks jms).

Kui tekstide masintõlkimisel ei ole dialoogile enamasti pööratud erilist tähelepanu, siis seoses kõnetöötluse arenguga on püstitatud selliseid ülesandeid, mis eeldavad kindlasti ka dialoogi modelleerimist, näiteks "tõlkiva telefoni" loomise probleem.

Kui arvutiga saab juba suhelda inimesele kõige loomulikumal viisil – kõne abil, siis võiks see suhtlemine toimuda ka inimestevahelise suhtluse normide järgi. See püstitabki nn konversatsiooni-

* Käesolev töö on valminud Eesti Teadusfondi toetusel (grant nr 4467).

agendi modelleerimise probleemi. Konversatsioonigent on teatavat liiki intelligentne agent (Feldman, Yu 1999), st arvutiprogramm, mis suudab suhelda inimesega täisväärtusliku partnerina – loomulikus keeles ja inimestevahelise suhtluse reegleid järgides (Allen 1994).

Siiski tuleb mainida, et erinevad uurijad suhtuvad erinevalt sellesse, kuidas peaks arvuti suhtlema inimesega. Ühed (nende hulgas ka käesoleva artikli autorid) arvavad, et see suhtlus peaks võimalikult sarnanema inimestevahelise suhtlusega, teised aga on vastupidisel seisukohal – kuna arvuti pole inimene, siis ei hakka inimene arvutiga iialgi suhtlema täpselt nii nagu inimesega. Kompromissi pole selles küsimuses siiani saavutatud (Dybkjær 2000).

Käesolevas artiklis on meie eesmärgiks käsitleda niisuguse konversatsiooniagendi modelleerimist, mis võiks osaleda nn loomulikus dialoogis, st loomulikus keeles ja inimestevahelise suhtlemise reeglite alusel toimuvas dialoogis. Artikkel on vahekokkuvõtteks meie aastatepikkusest tööst selles valdkonnas.

2. Taust

2.1. Kõnedialoogsüsteemid

Dialoogikomponenti on inimese ja arvuti vahelises suhtluses vaja mitmel põhjusel. Sageli ei väljenda kasutaja oma nõudlust ühe lausena või ühe vooruna (*turn*), sest see oleks ebapraktiline. Kasutaja ootab süsteemilt osavõttu, nii et suhtlus saaks loomulikul viisil kulgeda läbi mitme vooru. Dialoogsüsteem peab hoolitsema ka selle eest, et identifitseerida kõnet ja parandada vigu.

Inimese ja arvuti vahelise dialoogi uurimine on ajalooliselt kulgenud kahes suunas: diskursuse analüüs ja konversatsioonianalüüs (Giachin 1996).

Diskursuse analüüs, mis lähtub kõneaktide uurimisest, vaatleb dialoogi kui ratsionaalset kooperatsiooni ja eeldab, et rääkija lausungid on korrektsed laused. Konversatsioonianalüüs uurib dialoogi kui sotsiaalset interaktsiooni, kus vaadeldakse ka selliseid nähtusi, nagu ladususe katkemine, järsk fookuse ümberlülitamine jms.

Kõnedialoogsüsteemi tuumaks on dialoogihaldur, mille ülesandeks on dialoogi juhtimine vastavalt dialoogi mudelile. Rakendatav strateegia võib asuda kahe äärmuse vahel: ühelt poolt kasutaja täielik vabadus näidata initsiatiivi ja teiselt poolt täielikult dialoogihalduri määratav dialoog. Esimesel juhul on dialoog loomulik, kuid

suureneb risk, et süsteem mõistab kasutajat valesti. Teisel juhul on analüüs hõlpsam, kuid dialoogid võivad olla pikad ja ebasõbralikud.

Õige strateegia valik sõltub rakendusstsenaariumist ja kõnetuvastuse tõrkekindlusest. Sobiva strateegia määramine on oluline probleem, sest sellest oleneb suhtluse edukus. Hea strateegia on paindlik ja jätab kasutajale initsiatiivi senikauaks, kuni ei teki probleeme. Probleemide ilmnemisel aga nõuab dialoogihaldur kasutajalt ümbersõnastamist või kasutab muid suhtlusmodaalsusi: isoleeritud sõnad, tähthaaval hääldamine, jah-ei kinnitused.

Efektiivse kõnedialoogsüsteemi arendamine nõuab laiaulatuslikku eksperimenteerimist reaalsete kasutajatega. Süsteemi loomise algfaasis võib kasutada näiteks Võlur Ozi tehnikat: arvutit jäljendab inimekspert, kuid kasutajale jäetakse mulje, nagu ta suhtleks masinaga. Teine võimalik lähenemisviis on reaalsete kasutajatega eksperimenteerimine süsteemi loomise erinevatel etappidel. Mõlemat lähenemisviisi saab omavahel kombineerida. Igal juhul on aga dialoogsüsteemi arendamisel ja testimisel vaja ulatuslikku korpust reaalsetest või järeleaimatud dialoogidest.

Dialoogikorpust on vaja kas või ainult sellepärast, et kõneaktid võivad olla mitmefunktsionaalsed: ettepanek, nõustumine, tagasilükkamine jms ei tarvitse olla selgelt identifitseeritavad. Korpus annab empiirilist materjali nii teooriate arendamiseks kui ka reaalsete dialoogsüsteemide ülesehitamiseks. Seejuures erinevad kirjutatud dialoogid kõneldud dialoogidest, nii et dialoogikorpust koostama asudes tuleb endale täpselt selgeks teha, kas kasutada trükitud tekste või hankida materjali kõnesalvestustest.

2.2. Dialoogi mudelid

Dialoogi modelleerimisel seatakse eesmärgiks dialoogiteooria loomine ja algoritmide väljatöötamine, mis võimaldaksid arvutil osaleda dialoogis kasutajaga.

Praegu tuntakse kolme lähenemisviisi dialoogi modelleerimisele (Cohen 1996): 1) dialoogigrammatikad, 2) plaanipõhised meetodid, 3) ühistegevuse teooriad.

Dialoogigrammatikaid on arendatud alates 1970. aastate keskpaigast. See lähenemisviis põhineb tähelepanekul, et dialoogis esineb teatud arv järjestikusi regulaarsusi: näiteks küsimusele järgneb tüüpiliselt vastus, ettepanekule selle vastuvõtmine või tagasilükkamine jne. Dialooge käsitletakse kui selliste aktide

järjendeid. Dialoogigrammatika esitatakse kas generatiivse grammatikana või olekuautomaadina. Grammatikareeglitega seatakse kitsendused aktsepteeritavatele dialoogidele, nii nagu fraasistruktuurigrammatika reeglid asetavad kitsendused grammatiliselt korrektsetele lausetele. Dialoogigrammatika terminaalseteks elementideks on tavaliselt illokutsionaarsete kõneaktide nimed, nt küsimus, vastus, ettepanek jne. Mitteterminaalid kirjeldavad kindlat tüüpi dialoogide erinevaid staadiume, nagu initsieerimine, reageerimine, evalveerimine. Nii nagu fraasistruktuurigrammatika reegleid saab kasutada lausete analüüsimiseks, võib dialoogigrammatika reegleid kasutada dialoogi struktuuri analüüsimiseks.

Dialoogigrammatikad väljendavad enamasti ainult dialoogikäitumise lihtsaid seaduspärasusi. See mudel eeldab tüüpiliselt, et ühest olekust ülemineku tulemusel saabub üks kindel järgmine olek. Aga lausungid võivad olla mitmefunktsionaalsed: näiteks võib lausung olla ühekorraga nii keeldumine kui ka väitmine ning seega võib rääkija oodata kuulajalt vastust enam kui ühele tõlgendusele.

Plaanipõhised dialoogimudelid põhinevad tähelepanekul, et lausungid pole lihtsalt sõnade järjendid, vaid suhtlustegevuste, kõneaktide realiseerimine, nt ettepaneku esitamine, informeerimine, hoiatamine. Inimesed ei täida tegevusi juhuslikult, vaid planeerivad neid, et saavutada mitmesuguseid eesmärke, ning suhtlustegevuste korral on eesmärgiks ka muudatused kuulajate mentaalses seisundis. Näiteks kui rääkija planeerib käsku, siis on tema eesmärgiks ühtlasi muuta adressaadi kavatsusi. Suhtlustegevuse ja dialoogi plaanipõhised teooriad eeldavad, et rääkija kõneaktid on plaani elluviimise osa ning kuulaja ülesanne on tuvastada plaan ja vastata selle plaaniga kooskõlas.

See teooria on üldisem kui dialoogigrammatikad, kuna vaatleb dialoogi kui ratsionaalse mittekommunikatiivse käitumise erijuhtu. Plaanipõhises käsitluses on oluliseks komponendiks arvepidamine planeerimise ja plaani tuvastamise üle, rakendatakse mitmesuguseid tuletusreegleid ja tegevuse definitsioone, osalejate mentaalse seisundi mudeleid ning ootusi tõenäoliste eesmärkide ja tegevuste kohta kontekstis. Tegevusteks on kõneaktid, mille täitmine mõjutab dialoogis osalejate arvamusi, eesmärke, kohustusi ja kavatsusi. Selline kooperatiivse dialoogi mudel lahendab ka kaudsete kõneaktide probleemi. Näiteks kui rääkija ütleb: "Meil peaks kuskil olema nuga" siis peab adressaadi plaanituvastusprotsess määrama nii seda,

et rääkija tahab panna teda uskuma, et selline objekt eksisteerib, kui ka seda leidma ja rääkijale kätte toimetama.

Siiski on ka plaanipõhisel lähenemisel puudused, eelkõige teoreetilise baasi nõrkus. Näiteks on raske täpselt defineerida selliseid mõisteid, nagu plaan, eesmärk, kavatsus, ja kirjeldada dialoogis osaleja mentaalset seisundit.

Plaanipõhine lähenemine, mis modelleerib dialoogi lihtsalt kui sünkroonselt ja kooskõlastatult töötavate komponentide – plaanigeneraatorite ja -tuvastajate töö produkti, ei selgita, miks osalejad esitavad täpsustavaid küsimusi ja kinnitusi. Uus teooria – **ühistegevuse mudel** – käsitleb dialoogi kui midagi sellist, mida osalejad teevad üheskoos, eeldades, et kõik dialoogis osalejad vastutavad dialoogi kestmise eest.

Et osaleda dialoogis, peab partneritel olema vähemalt ühine kohustus üksteist mõista. See põhjendabki, miks inimestevahelises dialoogis on nii tavalised igasugused täpsustavad küsimused, selgitused, kinnitused.

Tüüpilised probleemid, mille lahendamisel sellised mudelid erinevad individuaalsetest plaanipõhistest mudelitest, on viitamiste (sh asesõnaline anafoor, deiksis, definiitsed ja deiktilised nimisõnafraasid) ja kinnituste käsitlemine. Tegelikku viitamiskäitumist ei saaks adekvaatselt modelleerida, kui lihtsalt kujutleda, et rääkija esitab nimisõnafraase ja kuulaja identifitseerib nende referente. Pigem esitavad mõlemad osalejad nimisõnafraase, täpsustavad eelmisi, parandavad valesid identifitseerimisi jne. See põhjendab oletust, et inimesed järgivad viitamisel nähtavasti ühise jõupingutuse minimiseerimise printsiipi.

Ühistegevuse teoreetilised mudelid püüavad minimiseerida dialoogis osaleva "meeskonna" kogujõupingutust dünaamilises, ebakindlas maailmas. Kui dialoogi käsitlemisel rakendada ühistegevuse teooriat, siis saab selgitada paljusid dialooginähtusi, nagu koostöö viitamisel, kinnitused jms. Selline teooria võimaldab kirjeldada, mida dialoogis osalejad peaksid tegema, et modelleerida seda käitumist kõnesüsteemides dialoogi juhtimisel. Edasine töö selles valdkonnas võib olla intelligentsete tarkvaraagentide suhtlusprotokollide aluseks.

2.3. Ratsionaalsete agentide mudelid

Diskursuse ja dialoogi pragmaatika lähtub eeldusest, et rääkija ja kuulaja on ratsionaalsed agendid. Ratsionaalsete agentide teooria käsitleb diskursust ja dialoogi kui käitumist, mis tuleneb agentide arvamustest, soovidest ja kavatsustest ja väljendab neid, olles samas piiratud agentide käsutuses olevate ressurssidega (Webber 2000). Nii planeerimine, st protsess, milles rääkija kavatsustele seatakse vastavusse tegevused, kui ka plaani tuvastamine, st protsess, mille abil kuulaja tuvastab rääkija kavatsusi, on otsinguprotsessid, kus tehakse järeldusi piiratud ressursside tingimustes.

Seejuures võivad rääkija ja kuulaja arvamused ja kavatsused dialoogi käigus muutuda.

Põhiline lähenemisviis planeerimisele lähtub eeldusest, et eesmärk (kavatsus) tekitab plaani; see on alguse saanud intellektitehnikast (Newell, Simon 1963). Kõige laiemalt on tuntud STRIPSi algoritm (Fikes, Nilsson 1971). Selle algoritmi andmestruktuurid väljendavad selliseid kavatsuste ja tegevuste tunnuseid, nagu asjaolu, et tegevustel on eeltingimused, mis peavad kehtima selleks, et tegevus saaks aset leida; need eeltingimused võivad omakorda olla teiste kommunikatiivsete tegevuste eesmärkideks; tegevustel on tulemused. Algoritmi järgmistesse versioonidesse on lisatud uusi võimalusi, näiteks käsitletakse tegevust mitmetasemelisena, vaadeldakse tegevusi info hankimisel, mis omakorda võib mõjutada tulevast plaani, arvestatakse asjaolu, et kui agendi arvamused muutuvad, siis võib ta püstitada uue eesmärgi.

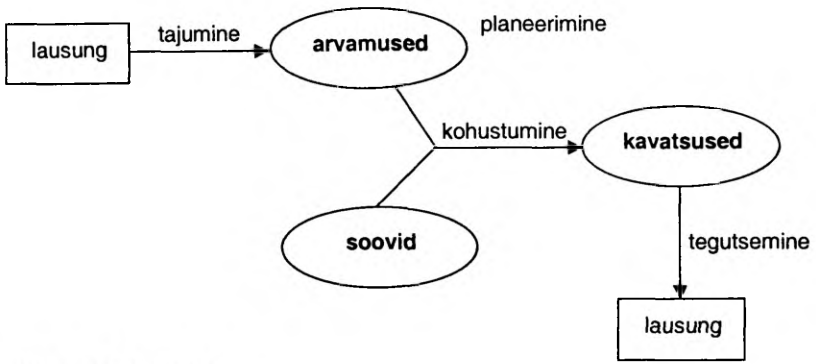
Praegu arendatakse keelelise suhtlemise kui ratsionaalselt planeeritava tegevuse käsitlemiseks ka keerulisemaid mudeleid, mis muuhulgas arvestavad järgmisi asjaolusid (Webber 2000).

1. Planeerija/rääkija arvamused võivad erineda kuulaja omadest ja olla ebakorrektsed.
2. Dialoogi ei saa kasutada ainult tegevuse lõpuleviimiseks, vaid ka tegevuse sooritamise erinevate võimaluste läbiuurimiseks ja kooskõlastamiseks.
3. Dialoog eeldab osalejate koostööd. Planeerimine on dialoogis keerukam kui üheainsa agendi puhul.
4. Planeerivatel agentidel on eelistused, mistõttu eesmärkide saavutamise ja tegevusplaanide realiseerimise viisid võivad olla erinevad.

5. Rääkija suhtlusakt ei tarvitse alati saavutada eesmärki, seetõttu peab rääkija kasutama tagasisidet kuulajalt, et vajaduse korral anda selgitusi.
6. Suhtlemisel antakse edasi informatsiooni, eesmärgiga realiseerida konkreetseid kavatsusi. Informatsiooni ja kavatsuste kombineerimist kasutatakse nii kommunikatiivsete tegevuste planeerimisel kui ka nende analüüsil.

2.4. Arvamuste, soovide ja kavatsuste modelleerimine

Inimestel on eesmärgid, mille saavutamiseks nad tegutsevad. Nad on teadlikud kehtivast olukorrast ja neil on positiivsed või negatiivsed tundmused selle olukorra suhtes. Sageli tegutsevad inimesed teataval viisil just selleks, et jõuda paremasse olukorda. Seega tuleb konversatsiooniagendi mudelis esitada nii agendi kognitiivset seisundit kui ka temas toimuvaid protsesse. Joonisel 1 on toodud üks selline mudel – arvamuse, soovi ja kavatsuse (*belief, desire, intention* ehk BDI) mudel (Allen 1994).



Joonis 1. BDI-mudel

Konversatsiooniagent, suheldes teise agendiga loomulikus keeles, tajub partneri lausungeid. Tal on arvamused kehtiva olukorra kohta ja soovid, mis võivad olla aluseks selle olukorra muutmisel. Arvamused saavad agendis käivitada planeerimisprotsessi ning koos soovidega kohustuse võtmise (*commitment*) protsessi, mille tulemuseks on kavatsused teatavaid plaane ellu viia. Kavatsused omakorda käivitavad agendi tegutsemise, mille tulemusel ta genereerib vastuse partnerile – lausungid. Suhtlemise käigus korduvad need protsessid

tsükliliselt. See mudel käsitleb lihtsustatud konversatsiooniagenti, mis tajub ainult lausungeid ja sooritab ainult lausungite genereerimise tegevusi.

2.5. Kooperatiivse probleemilahendusdialoogi modelleerimine

Huvi koostöödialoogi vastu, kus osalejad tegutsevad ühise eesmärgi nimel, on taas tõusnud seoses Interneti-teenuste laia levikuga. Enamasti käsitletakse kooperatiivseid probleemilahendusdialooge.

Üldtuntud probleemilahendusmudeliks on intelligentne ressursiseotud masinarhitektuur – *Intelligent Resource-bounded Machine Architecture* ehk IRMA (Bratman 1988), mis kujutab endast BDI-mudeli ühte realisatsiooni. Ressursiseotus tähendab siin seda, et fikseeritud aja jooksul ei saa agent teha suvalise pikkusega arvutusi. IRMA kasutab probleemilahendusel vahendianalüüsi (*means-end analysis*). Aluseks on võetud idee, et agendi kavatsused, mille alusel ta on koostanud plaanid, kitsendavad ja suunavad arutlust, esiteks vahendianalüüsi rakendamisel ja teiseks tegevuse valikul võimalike alternatiivide hulgast.

Seda arhitektuuri on täiendanud Clark, lisades sellesse kooperatiivsuse. Clarki mudeli aluseks on eeldus, et rääkimine ja kuulamine pole autonoomsed tegevused, vaid kollektiivse tegevuse osad. Ta eristab nelja taset:

1) A teeb tegevusi ja B jälgib neid;
2) A annab signaali ja B identifitseerib selle;
3) A signaliseerib B-le midagi ja B tuvastab A mõtte;
4) A teeb ettepaneku ühisprojekti elluviimiseks ja B kas aktsepteerib seda või lükkab tagasi.

Paljud uurijad on modelleerinud nõustumisprotsessi kooperatiivses dialoogis, so. situatsiooni, kus üks osaleja, A, teeb partnerile B ettepaneku ja B kas aktsepteerib seda või mitte. Enamus uurimusi keskendub Clarki 4. tasemele, st ettepanekutele ja nende aktsepteerimisele või tagasilükkamisele.

Chu-Carroll ja Carberry (1998) esitavad kooperatiivse vastuste genereerimise mudeli kui rekursiivse tsükli ettepanek–hindamine–modifitseerimine (*Propose–Evaluate–Modify*). Nad keskenduvad infojagamis- ja läbirääkimisdialoogidele. Infojagamisdialoog algatatakse, kui agent on tuvastanud partneri tehtud ettepaneku, kuid tal pole piisavalt infot, et see vastu võtta või tagasi lükata.

Läbirääkimisdialoog algatatakse, kui agent tuvastab, et ettepanek on konfliktis tema arvamustega, st kaldub seda tagasi lükkama.

Heeman ja Hirst (1995) modelleerivad koostööd tsükli esitamine–arvustamine–ümberkujundamine (*Present–Judge–Refashion*) abil. Nad kasutavad kahte taset: planeerimine ja koostöö. Esimesel tasemel interpreteeritakse ja genereeritakse lausungeid, teisel tasemel aga modelleeritakse agentide koostööd, seostades seda kummagi agendi mentaalse seisundi ja planeerimisprotsessidega.

Ühistegevuse mudel SharedPlans (Lochbaum 1998) käsitleb planeerimisprotsessi, milles osalevad mitu agenti. Kesksel kohal on agentide kavatsuste tuvastamine ja nende koordineerimine ühise eesmärgi saavutamiseks. Keskendutakse grupiülesannetele, mida saab lahutada eraldiseisvateks, kuid kooskõlastatud individuaalseteks plaanideks.

Di Eugenio jt (2000) esitavad mudeli kaalumine–ettepanek–seisukohavõtt (*BalanceProposeDispose*): esmalt kaalutakse infot, arutatakse, peetakse nõu, siis tehakse ettepanek ja lõpuks leitakse ettepaneku koht – teda kas aktsepteeritakse või lükatakse tagasi.

2.6. Dialoogi struktuur

Nagu varem mainitud (2.1), võetakse dialoogiteooria arendamisel või dialoogsüsteemi loomisel sageli aluseks dialoogikorpus, milles on märgendatud need nähtused, mis konkreetset uurijat huvitavad. Näiteks Carletta jt (1997) kasutavad 3-tasemelist kodeerimissüsteemi. Nende dialoogikorpus MapTask koosneb dialoogidest, kus kummalgi osalejal on pisut erinev versioon ühest ja samast kaardist ja teine osaleja peab suutma oma kaardile märkida sellesama tee, mis algselt on kujutatud partneri kaardil. Osalejad asuvad erinevates ruumides ja suhtlevad arvuti abil.

Kõige kõrgemal tasemel jaotatakse dialoog **transaktsioonideks** (*transaction*). Transaktsiooni all mõistetakse ühe ülesande lahendamisel tekkivat alamdialoogi. Korpuses MapTask loetakse tüüpiliseks transaktsiooniks alamdialoogi, mille tulemusel teine osaleja on mõistatanud ühe lõigu teest.

Transaktsioonid koosnevad omakorda konversatsioonimängudest ehk dialoogimängudest ehk **vahetustest** (*exchange*). Vahetus on lausungite järjend, mis algab initsieerimisega ja lõpeb eesmärgi saavutamise või sellest loobumisega, näiteks küsimus, millele järgneb vastus, või ettepanek, millele järgneb selle vastuvõtmine või

sellest loobumine. Vahetus koosneb vähemalt kahest erineva kõneleja **voorust** (*turn*).

Voor omakorda on kõik see, mida kõneleja ütleb enne, kui järgmine jutu üle võtab (Stenström 1994). Vooru pole muidugi korpuses vaja märgendada, sest ta on füüsiliselt tuvastatav. Iga voor võib koosneda ühest või mitmest **sammust** (*move*). Korpuses MapTask liigitatakse samme initsieerimisteks (nt käsk, selgitus, küsimus) ja vastamisteks (nt kättesaamisteade, selgitus, jaatus). Iga samm võib koosneda ühest või mitmest **kõneaktist**. Kõneakt on väikseim interaktsiooni ühik. Ta näitab, mida kõneleja kavatseb, milleks ta tahab suhelda (Stenström 1994). Korpuses MapTask esinevad üksnes ühest aktist koosnevad sammud, mistõttu aktide ja sammude vahel pole vahet tehtud.

Toome veel mõned näited dialoogide märgendamise kohta. Walker jt (1990) märgendavad väiteid, käske, küsimusi, viipu (*prompt*), eesmärgiga uurida initsiatiivi vaheldumist. Sutton jt (1995) märgendavad ainult vastuseid (st kas vastati küsimusele või mitte). Projekti VERBMOBIL dialoogikorpus (Alexandersson jt 1995) sisaldab ärikohtumiste aja kokkuleppimise dialooge, mille märgendamisel on kasutatud 17 kõneakti. Analoogilises korpuses kasutavad Nagata jt (1993) 9 kõneakti. Ahrenberg jt (1995) uurivad dialoogi fookuse struktuuri infootsidialoogides, mis on genereeritud Võlur Ozi tehnikat kasutades, ja märgendavad samme.

Eestikeelset dialoogikorpust praegu veel ei eksisteeri.

3. Üks konversatsiooniagendi mudel

Dialoogi võib käsitleda kahel viisil: ühelt poolt kui suhtlusprotsessi, milles osalejad püüavad saavutada oma kommunikatiivseid eesmärke, viies läbi arutlusi plaanide koostamiseks ja rakendades suhtlusstrateegiaid plaanide elluviimiseks, ning teiselt poolt kui teksti või kõnet, mis on selle protsessi materialiseerunud jälg.

Mitmes oma varasemas artiklis oleme välja töötanud dialoogi (protseduraalse) mudeli (Koit, Õim 1993, 1994, 2000, Koit 1996, Õim 1996). Selle mudeli kohaselt võib suhtlusprotsessis osalejat ehk konversatsiooniagenti A esitada kui 6 komponendist (moodulist) koosnevat programmi

A = (PL, ÜL, DH, INT, GEN, KP)

PL – planeerija, ÜL – ülesannete lahendaja, DH – dialoogihaldur,
INT – interpretaator, GEN – generaator, KP – keeleprotsessor

Planeerija on "keskprotsessor" mis annab korraldusi nii dialoogihaldurile kui ka probleemilahendajale, seejuures dialoogihaldur juhib suhtlust ja probleemilahendaja lahendab ainevaldkonna ülesandeid. Keeleprotsessor teeb partneri lausungi morfoloogilist ja süntaktilist analüüsi ning agendi enda lausungite süntaktilist ja morfoloogilist sünteesi. Interpretaatori ülesandeks on partneri lausungite semantiline analüüs ja generaatori ülesandeks agendi enda lausungite semantiline süntees.

Konversatsiooniagent kasutab oma töös eesmärkide baasi EB ja teadmusbaasi TB.

Meie mudeli kohaselt koosneb teadmusbaas omakorda 4 komponendist:

$TB = (TB_M, TB_K, TB_D, TB_S)$

TB_M – teadmus ainevaldkonna (maailma) kohta, TB_K – keeleteadmus, TB_D – dialoogiteadmus, TB_S – teadmus isikute e subjektide (teiste agentide ja iseenda) kohta

Teadmus ainevaldkonna kohta sisaldab ainevaldkonna objektide ja nendevaheliste suhete definitsioonid (deklaratiivne teadmus) ja ülesannete lahendamise algoritmid (protseduraalne teadmus). Keeleteadmus koosneb kasutatava(te) keel(t)e leksikonidest (deklaratiivne teadmus) ning teksti ja/või kõne analüüsi ja sünteesi algoritmidest (protseduraalne teadmus). Dialoogiteadmus sisaldab suhtlussammude, voorude, vahetuste ja transaktsioonide definitsioonid ning suhtluseesmärkide saavutamiseks rakendatavad algoritmid, mida me nimetame suhtlusstrateegiateks ja -taktikateks. Teadmus isiku(te) kohta koosneb, ühelt poolt, nende (tegelikest või oletatavatest) hinnangutest maailmale (nt mida nad peavad meeldivaks või ebameeldivaks, kasulikuks või kahjulikuks) ja teiselt poolt, algoritmidest, mida nad rakendavad, et hinnangute alusel genereerida tegevusplaane.

Tarvilik tingimus selleks, et suhtlus saaks üldse aset leida, on jagatud teadmuse olemasolu: kõigil osalejail peab olema vähemalt ühine eesmärk suhelda, ühine keel, ühine ettekujutus ainevaldkonnast ja suhtlemisnormidest ning vähemalt osaliselt õige ettekujutus suhtluspartnerist. Kui dialoogis osalejad on S_1 ja S_2, siis jagatud teadmus tähendab, et

$EB_1 \cap EB_2 \neq \emptyset$, $TB_{K1} \cap TB_{K2} \neq \emptyset$, $TB_{M1} \cap TB_{M2} \neq \emptyset$, $TB_{D1} \cap TB_{D2} \neq \emptyset$, $TB_{S12} \cap TB_{S2} \neq \emptyset$ ja $TB_{S21} \cap TB_{S1} \neq \emptyset$.

Keeleprotsessor teeb partneri lausungi morfoloogilise ja süntaktilise analüüsi, saadud süntaktiline esitus läheb järgmisse plokki – interpretaatorisse. See teeb lausungi semantilise analüüsi ja määrab ka tema kommunikatiivse struktuuri, st tuvastab, kas on tegu nt küsimuse, ettepaneku, vastuse või muu kõneaktiga (suhtlussammuga). Ühtlasi püstitab interpretaator kaht liiki eesmärgid. Esiteks, kommunikatiivne ehk suhtluseesmärk sõltub repliigi kommunikatiivsest struktuurist: näiteks ettepanekule järgneb tüüpiliselt selle vastuvõtmine või tagasilükkamine, küsimusele vastus jne. Suhtluseesmärk ongi seega vastata ettepanekule, vastata küsimusele, esitada küsimus teatava info hankimiseks jne. Teiseks, ainealane eesmärk sõltub repliigi semantilisest struktuurist ja seisneb konkreetse ülesande lahendamises. Nt vastamaks küsimusele "Kui palju maksab lennukipilet Tallinnast Amsterdami läbi Helsingi?" tuleb see teadaolevatest piletihindadest arvutada.

Esimest tüüpi eesmärkide töötlemisega tegeleb dialoogihaldur, teist tüüpi eesmärkidega aga ülesannete lahendaja. Dialoogihalduri töö tulemusel määratakse vastuslausungi kommunikatiivne struktuur (st otsustatakse, missugustest suhtlussammudest koosneb lausung), ülesannete lahendaja töö tulemusel aga vastuslausungi sisu.

Generaator moodustab vastuslausungi semantilise esituse ning lingvistiline protsessor vormistab selle loomulikus keeles.

Käsitleme järgnevas suhtlemise erijuhtu, kus osaleja A kommunikatiivseks eesmärgiks on saavutada partnerilt B nõusolek sooritada teatavat tegevust D. Suhtlusprotsessi võib vaadelda nii A kui ka B seisukohalt.

3.1. Arutluse modelleerimine

Dialoogis osaleja ei vali järgmist kõneakti juhuslikult, vaid enne valiku tegemist viib läbi arutluse. Kui A on esitanud B-le ettepaneku teha tegevus D, siis B vastab nõustumise või keeldumisega, sõltuvalt sellest, missugune on tema arutluse tulemus. Meie poolt valitud suhtluse erijuhu üks väärtusi ongi selle suhteliselt selge piiritletus: arutluse sisuks on B hinnangud tegevuse D teatud aspektidele ja nende hinnangute võrdlemine – mis kaalub mille üles?

Kuna meie arvates on arutlusmudel dialoogi loomulikkuse saavutamise oluline komponent, siis selgitame järgnevas selle aluseks olevaid printsiipe (vt ka Koit, Õim 1994; Õim 1996).

Üldisemas mõttes on siin tegu eespool juba mainitud BDI-mudeliga. Oleme püüdnud konkretiseerida eelkõige selle kaht komponenti: arvamused ja soovid. Meie arutlusmudel sisaldab kaks omavahel funktsionaalselt seotud osa. Esiteks, inimese motivatsioonisfääri mudeli ja teiseks, arutlusalgoritmid. Tuleb rõhutada, et me ei järgi siin ühtegi inimmotivatsiooni ja -arutluse teaduslikku teooriat, vaid modelleerime "naiivset arutlusteooriat" mida meie arvates inimesed järgivad intuitiivselt, kui nad suhtlevad teiste inimestega, püüdes mõista, ennustada või mõjutada oma partneri otsuseid ja tegevusi (Õim 1996).

Meie mudeli kohaselt reguleerivad inimese arutlust selle üle, kas teha või tegemata jätta tegevus D, kolme liiki faktorid (determinandid), mis omakorda jagunevad arutleva subjekti seisukohalt sisemisteks ja välimisteks. Sisemised faktorid on subjekti soovid ja vajadused, välimised aga tema kohustused.

Subjekt *soovib* teha tegevust D, kui tema jaoks ületavad D meeldivad aspektid ebameeldivaid. Subjektil on *vaja* teha D, kui kasu D tegemisest ületab kahju. Subjekt on *kohustatud* tegema D, kui D tegematajätmine toob kaasa karistuse (mis on ju ebameeldiv või kahjulik). Püüdes jõuda otsusele D tegemise suhtes, lähtub arutlev subjekt esmalt oma soovist, st kontrollib, kas D meeldivad aspektid ületavad ebameeldivaid. Kui see on nii, siis kontrollib subjekt ressursside olemasolu ning kui need on olemas, siis kaalub D muid positiivseid ja negatiivseid külgi: kasulikkust ja kahjulikkust, ning kui D on keelatud tegevus, siis ka selle tegemisele järgnevat karistust. Kui positiivsed aspektid summaarselt ületavad negatiivseid, siis on arutluse tulemuseks otsus teha D, vastupidisel juhul aga tegemata jätta.

Kuna selle mudeli kohaselt tuleb summeerida tegevuse erinevate aspektide (meeldivus, ebameeldivus, kasulikkus, kahjulikkus, karistus keelatud tegevuse tegemise või kohustusliku tegevuse tegematajätmise eest) kaalusid, siis järelikult peavad neil kaaludel olema arvulised väärtused. Tegelikult inimesed muidugi ei opereeri sellises arutlusprotsessis arvudega, vaid pigem hägusate hulkadega. Näiteks võrreldes tegevuse meeldivaid ja ebameeldivaid külgi, kasutatakse selliseid sõnu nagu *suurepäranev, tore, vastuvõetav, mitterahuldav, vastik* jms. Iga sellise adjektiivi võib aga taandada teatavale arvulisele skaalale, saades sel teel vaadeldava tegevuse vastava aspekti kaalu arvulise väärtuse.

Agendi A motivatsioonisfääri mudeliks oleme siin võtnud kaalude vektori. (Arutlusmudeli teine osa – arutlusalgoritmid – on esitatud artiklis Koit 1996.)

k^A = (k(ressursid$^A_{D1}$), k(meeldiv$^A_{D1}$), k(ebameeldiv$^A_{D1}$), k(kasulik$^A_{D1}$), k(kahjulik$^A_{D1}$), k(kohustuslik$^A_{D1}$), k(keelatud$^A_{D1}$), k(karistus$^A_{D1}$), k(karistus$^A_{mitte\text{-}D1}$), ..., k(ressursid$^A_{Dn}$), k(meeldiv$^A_{Dn}$), k(ebameeldi$^A_{Dn}$), k(kasulik$^A_{Dn}$), k(kahjulik$^A_{Dn}$), k(kohustuslik$^A_{Dn}$), k(keelatud$^A_{Dn}$), k(karistus$^A_{Dn}$), k(karistus$^A_{mitte\text{-}Dn}$))

D_1, ..., D_n – (kõikvõimalikud) inimtegevused; k(ressursid$^A_{Di}$) = 1, kui A omab kõik vajalikud ressursid tegevuse D_i tegemiseks (vastasel korral 0); k(kohustuslik$^A_{Di}$) = 1, kui D_i on A jaoks kohustuslik (vastasel korral 0); k(keelatud$^A_{Di}$) = 1, kui D_i tegemine on A jaoks keelatud (vastasel korral 0). Ülejäänud kaalude väärtusteks on mittenegatiivsed täisarvud.

Arutlusmudel on konversatsiooniagendi A mudeliga seotud sel teel, et esiteks, planeerija kasutab arutlusalgoritme ja teiseks, teadmusbaasi osa TB_S sisaldab nii kaalude vektori k^A (A isiklikud hinnangud kõikvõimalikele tegevustele) kui ka vektorid k^{AB} (A arvamused B hinnangute kohta, kus B on kõikvõimalikud teised subjektid, kellega A saab suhelda). Viimased vektorid ei esita muidugi tõsikindlat teadmust, vaid selle usaldusmäär on väiksem kui 1. Vektoreid k^{AB} kasutame partneri(te) mudelina.

Kui tõmmata nüüd paralleele BDI-mudeliga, siis on konversatsiooniagendi *arvamusteks* teadmised, usaldusmääraga alla 1; *soovid* genereeritakse kaalude vektoriga k^A ning *kavatsused* on eesmärgid (eesmärkide) baasist EB. Lisaks soovidele on kaalude vektorist tuletatavad ka agendi motivatsioonisfääri mõned sellised parameetrid, mida BDI-mudel ei hõlma: vajadused, kohustused ja keelud, kusjuures ühed soovid (või vajadused) võivad olla suuremad kui teised (näiteks kui k(meeldiv$^A_{Di}$) > k(meeldiv$^A_{Dj}$), siis on soov teha tegevust D_i suurem kui soov teha tegevust D_j) ja ühed kohustused (või keelud) võivad olla rangemad kui teised (sõltudes karistuse suurusest).

3.2. Suhtlusstrateegiad ja -taktikad

Dialoogiteadmus TB_D, mida kasutab konversatsiooniagendi dialoogihaldur DH, koosneb ühelt poolt suhtlemise reeglitest ja teiselt poolt kõneaktide konstrueerimise ning omavahel kombineerimise reeglitest.

Suhtlemise reeglid võtab kokku suhtlusstrateegia – algoritm, mida agent kasutab oma suhtluseesmärgi saavutamiseks. Artiklis (Koit 1996) on esitatud algoritm, mida A rakendab, saavutamaks partneri B otsust teha tegevus D.

Agent võib suhtlusstrateegiat realiseerida erinevate suhtlustaktikate kaudu. Näiteks saab A ahvatleda, veenda või ähvardada partnerit B tegema tegevust D (vt Koit 1996). Ahvatlemise puhul rõhutab A tegevuse meeldivust, veenmise puhul kasulikkust ja ähvardamise puhul (kohustusliku) tegevuse tegematajätmisele järgnevat karistust. Missuguse suhtlustaktika A valib, sõltub sellest, kas suhtlus on kooperatiivne või hoopis konfrontaalne (näiteks tülitsemine), isiklik või isikupäratu (näiteks ametikõnelus), missugune on suhtlusdistants suhtlejate vahel (näiteks kas nad on sõbrad või hoopis ülemus ja alluv), missugune on suhtlemise modaalsus (näiteks sõbralik või ebasõbralik) ja intensiivsus (vaoshoitud või keevaline). Neid suhtluse parameetreid nimetame suhtlusruumi koordinaatideks ja nende väärtused on iseloomustatavad vastavate adjektiividega, mida meie oma mudelis taandame arvulistele skaaladele nagu eespool käsitletud tegevuse aspektide kaalude väärtusigi (Koit, Õim 1993; Õim, Koit 1994). Teisisõnu: suhtlustaktika valik sõltub sellest, missuguses suhtlusruumi punktis osalejad parajasti asuvad.

3.3. Kõneaktid

Suhtlemisel sooritavad konversatsiooniagendid kõneakte. Kõneakt on vähim interaktsiooni ühik. Meie kasutame oma mudelis piiratud kogust kõneakte, mille esitusformalismiks oleme valinud freimid.

Iga kõneakt sisaldab staatilise ja dünaamilise osa. Staatiline (deklaratiivne) osa koosneb 1) eeltingimustest, 2) eesmärgist, 3) sisust ja 4) tulemustest. Dünaamiline (protseduraalne) osa sisaldab kaht liiki protseduure:

1) need, mida kõneakti autor rakendab vaadeldavat kõneakti sisaldava suhtlussammu genereerimiseks;
2) need, mida adressaat rakendab selle kõneakti interpreteerimiseks ja oma vastuse genereerimiseks. Toome näiteks kõneakti "ettepanek" freimi, kus on eeldatud kooperatiivset suhtlust osalejate vahel.

ETTEPANEK
(autor A, adressaat B: A teeb B-le ettepaneku teha tegevus D)

I. Staatiline osa
EELTINGIMUSED:
(1) A-l on eesmärk E
(2) A arvab, et ka B-l on eesmärk E
(3) A arvab, et E saavutamiseks on vaja esmalt saavutada vahe-eesmärk Ev
(4) A arvab, et ka B arvab, et E saavutamiseks on vaja esmalt saavutada vahe-eesmärk Ev
(5) A arvab, et Ev saavutamiseks peab B tegema D
(6) A arvab, et B omab ressursid D tegemiseks
(7) A arvab, et B otsustab teha D
EESMÄRK: B otsustab teha D
SISU: A teatab B-le, et B tehku D
TULEMUSED:
(1) B teab Eeltingimusi, Eesmärki ja Sisu
(2) A teab, et B teab Eeltingimusi, Eesmärki ja Sisu

II. Dünaamiline osa

Genereerimisprotseduurid (A võimalused ehitada oma repliiki, mis sisaldab ettepaneku)
A-l on eesmärk E; ta teab (eeldab), et ka B-l on eesmärk E; ta eeldab, et E saavutamiseks on vaja saavutada Ev; A on otsustanud vormistada selle ettepanekuna B-le.

Protseduurid (enne ettepaneku väljaütlemist) eeltingimuste kontrollimiseks:
(2) korral – kas B-l on aktualiseeritud E? Kui ei, siis aktualiseerida see teatamisega.
(4) korral – kas B arvab, et E saavutamiseks on vaja saavutada Ev? Kui ei, siis lisada ettepanekule seletus (argument) Ev vajalikkuse kohta (KUI Ev, SIIS E).
(6) korral – kui A kahtleb, kas B-l on ressursid D tegemiseks, siis lisada ettepanekule küsimus (kas ...?).
(7) korral – kui A pole kindel, kas B otsustab teha D (A peaks läbi mängima B võimalikud arutlused), siis lisada argument.

• *Interpreteerimis-genereerimisprotseduurid* (B võimalused reageerida ettepanekule)
Käivitatakse eeltingimuste kontrollil pärast ettepaneku tuvastamist.
(2), (4), (5) korral – kui B-l pole eesmärki E ja/või ta ei arva, et E saavutamiseks on vaja saavutada Ev ja selleks omakorda B-l teha D, siis küsimus (küsida lisainfot).
(6) korral – kui B-l pole ressursse, siis eitus+argument.
(7) korral – kui B otsus teha D on eitav, siis eitus+argument.

Selles freimis esinevad viited teistele kõneaktidele: argument, küsimus, teatamine (info andmine). Järgnev näide esitab freimi "argument"

ARGUMENT
(= põhjendamine: autor A põhjendab adressaadile B väitega X väidet Y)
I. Staatiline osa
EELTINGIMUSED:
(1) A arvab, et kehtib X
(2) A arvab, et kehtib Y
(3) A arvab, et kui kehtib X, siis kehtib Y
(4) A arvab, et B arvab, et kui kehtib X, siis kehtib Y
EESMÄRK: B arvab, et kehtib väide Y
SISU: A teatab B-le, et kehtib X
TULEMUSED:
(1) B arvab, et kehtib X
II. Dünaamiline osa
• *Genereerimisprotseduurid* (A võimalused ehitada oma repliiki, mis sisaldab argumendi):
teatada B-le X või teatada B-le, et kehtib X ja kui kehtib X, siis kehtib ka Y
• *Interpreteerimis-genereerimisprotseduurid* (B võimalused reageerida A argumendile):
aktsepteerimine või (vastu)argument

Selline kõneakti kaheosaline esitus garanteerib ühtlasi dialoogi (tegelikult dialoogimängu ehk vooruvahetuse) sidususe: kui dialoogiteadmiste baasis TB_D märgendada initsieerivad kõneaktid (nagu küsimus või ettepanek), siis järgnev kõneaktide ahel tuletub adressaatide interpreteerimis-genereerimisprotseduuride rakendamisest.

Üldisemal juhul peab TB_D aga sisaldama ka dialoogistsenaariumide graafi, mille sõlmedeks on erinevate osalejate kõneaktid ja servadeks võimalikud vahetud üleminekud.

3.5. Maailmateadmus

Ainevaldkonnateadmiste (maailmateadmuse TB_M) esitamise formalismina kasutame samuti freime.

Vaadeldaval erijuhul huvitavad meid eeskätt mitmesuguste tegevuste kirjeldused.

Tegevuse freim sisaldab sellised slotid nagu EELTINGIMUSED, EESMÄRK, TULEMUSED, AKT (tegevuse lahutus elementaartegevusteks), TEGIJA, KOHT, AEG jms.

Vaatleme näiteks situatsiooni, kus A teeb B-le ettepaneku valmistada kartulisalat (st tegevuseks on kartulisalati valmistamine). Selle tegevuse freim on üldise tegevusfreimi alamfreimiks.

Kartulisalati_valmistamine
ÜLEMFREIM: TEGEVUS
RESSURSID:
- Materjal: keedetud kartulid, kõvaks keedetud muna, hapukurk, hakitud sibulapealsed, hapukoor, sool
- Vahend: kauss
- Ajakulu: 5–30 min
- Nõutavad oskused: võta, tükelda, sega, maitsesta, kaunista

AKT:
- võta keedetud kartulid, kõvaks keedetud muna, hapukurk, hakitud sibulapealsed, hapukoor, sool
- tükelda kartulid, muna ja kurk
- sega kausis koos hapukoorega
- maitsesta soolaga
- kaunista sibulapealsetega

EELTINGIMUSED: TEGIJA omab RESSURSID
EESMÄRK: kartulisalat
TULEMUS: kartulisalat

Selle freimi üks konkreetne eksemplar võib olla näiteks järgmine:
Kartulisalati_valmistamine_1
RESSURSID:
- Materjal: 4 keedetud kartulit, 1 kõvaks keedetud muna, 1 hapukurk, 1 spl hakitud sibulapealseid, 3 spl hapukoort, soola;
- Vahend: kauss mahuga >= 1 liiter
- Ajakulu: 15 min
- Oskused: võta, tükelda, sega, maitsesta, kaunista

TEGIJA: B
KOHT: A köök
AEG: täna
AKT:
- võta 4 keedetud kartulit, 1 kõvaks keedetud muna, 1 hapukurk, 1 spl hakitud sibulapealseid, 3 spl hapukoort, soola
- tükelda kartulid, muna ja kurk
- sega kausis koos hapukoorega
- maitsesta soolaga
- kaunista sibulapealsetega

EELTINGIMUSED: TEGIJA omab RESSURSID
EESMÄRK: kartulisalat
TULEMUS: kartulisalat

Siin eeldatakse, et A ja B asuvad A köögis ning B valmistab kartulisalati vastava tegevusfreimi ressursside osas loetletud materjalist, kulutades selleks 15 minutit.

3.6. Suhtlusprotsessi kulg

Dialoog kulgeb järgmiselt.

Suhtluspartnerid on A ja B. Meie juhul on A suhtluseesmärk 'B teeb D' Kirjeldame siin olukorda, kus nii A kui ka B on

intelligentsed agendid, st suhtlus toimub kahe arvutiprogrammi vahel.

1. A moodustab

a) tegevuse D freimi eksemplari, koondades sellesse kogu temal olemasoleva info tegevuse kohta;

b) partneri B mudeli, koondades sellesse kogu temal olemasoleva info partneri hinnangutest D freimi eksemplari slottidele.

2. A valib suhtluspunkti, millest ta peab konkreetse B puhul lähtuma.

3. A hakkab rakendama suhtlusstrateegiat. A mängib läbi B arutluse, kasutades arutlusalgoritmi ja B mudelit. Algul rakendab A soovist lähtuvat arutlust. Kui arutlus annab tulemuseks 'teha D', siis aktualiseerib A ahvatlemise taktika ja moodustab oma esimese lausungi, mis sisaldab ettepaneku teha D, st freimi ETTEPANEK eksemplari. Kui soovist lähtuv arutlus annab tulemuseks 'mitte teha D' siis proovib A vajadusest lähtuvat arutlust ja seejärel kohustusest lähtuvat arutlust, aktualiseerib kas veenmise või ähvardamise taktika ja moodustab esimese repliigi. Kui iga arutlus annab tulemuseks 'mitte teha D', siis vastavalt suhtlusstrateegiale valib A (edutu) lõpetamise (suhtlus jääb toimumata).

4. B interpreteerib A lausungit ja tuvastab selles ettepaneku. B moodustab omakorda tegevuse D eksemplari (mis ei tarvitse ühtida A omaga). Käivitab arutluse, mille käigus peab võibolla A-lt lisainfot küsima (ja seda saama). Vastavalt suhtlussammu ETTEPANEK freimile väljastab B arutluse tulemuse (jah/ei + võibolla argument).

5. A interpreteerib B vastust ja tuvastab, millises dialoogistsenaariumi punktis see asub. Kui B vastus oli jaatav (otsus teha D), on vastavalt suhtlusstrateegiale saabunud edukas lõpp. Kui B vastus oli eitav, siis dialoogistsenaariumi kohaselt peab A moodustama (vastu)argumendi. Suhtlustrateegia näeb ette ka võimaluse suhtluspunkti või -taktikat muuta. Vastuargumendi moodustamiseks võtab A infot D eksemplarist (mida ta võibolla täiendas B keeldumisest saadud infoga), arvestades endal olemasolevat B mudelit (mida ta pidi B keeldumise tõttu muutma). A mängib jälle läbi B arutluse, nagu dialoogi alustades, st kõik kordub tsükliliselt.

Nii A kui ka B hindavad tegevust selle tegija B seisukohast. Selle hindamise tulemusel saadakse isikumudelid – kaalude vektorid:
1) k^{B}on B (tegelikud) hinnangud tegevusele ehk B enda mudel;
2) k^{AB} on A ettekujutus B hinnangutest tegevusele (mis ei tarvitse olla õiged) ehk partneri mudel.

Hinnatakse, esiteks, ressursse – mida on, mida ei ole, mil määral on. Tegevuse sooritamiseks vajaminevad ressursid on loetletud tegevusfreimis. Ressursid on "objektiivsed" selles mõttes, et ei sõltu tegijast. Hinnangud ressursside olemasolule seevastu võivad muidugi olla subjektiivsed.

Mis on meeldiv või ebameeldiv, kasulik või kahjulik, kohustuslik või keelatud või mis on karistus – see sõltub konkreetsest subjektist.

Eraldi hinnatakse tegevuse D iga elementaartegevuse meeldivust, ebameeldivust jne ja koondhinnang arvutatakse kui nende hinnangute summa.

Vaatleme näidet võimalikust dialoogist: A ja B asuvad A köögis; A eeldab, et B omab kõik ressursid tegevuse tegemiseks ja peab tegevust pigem kasulikuks kui kahjulikuks; A *veenab* B-d valmistama kartulisalatit (st püüab saavutada olukorda, et B otsustaks valmistada salati, pidades seda tegevust kasulikuks).

A: Valmista kartulisalatit.	ettepanek
B: Pean koostama ettekannet homseks koosolekuks.	keeldumine argumendi esitamise teel (ressursse – aega – napib); A ettekujutus B-st oli vale => ressursid-1
A: Ma aitan sul ettekannet teha.	argument: ressursid on hangitavad => ressursid+1
B: Kodus hakatakse minu pärast muretsema.	argument: ressursse – aega – napib => ressursid-1
A: Helista ja ütle, et jääd kauemaks.	argument: ressursid on hangitavad => ressursid+1
B: Ma ei oska kartuleid tükeldada.	argument: ressursse napib => ressursid-1
A: Kutsu sõber appi.	argument: ressursid on hangitavad => ressursid+1
B: Köögis on alati väga palav.	argument: kahjulikkus suur => kahjulikkus +1
A: Minu köögis on hea ventilatsioon.	argument: selle vähendamine => kahjulikkus -1

B: Ma võin sõrme lõigata.	argument: kahjulikkus suur => kahjulikkus +1
A: Mul on kvaliteetsed abivahendid.	argument: selle vähendamine => kahjulikkus -1
jne	

Kui A või B rollis on inimene, siis ta muidugi ei opereeri freimidega ega kasuta eespool nimetatud algoritme, vaid suhtleb nagu inimene ikka. Arvuti – tema partner, konversatsiooniagent – toimib nii, nagu oleme eespool kirjeldanud.

4. Kokkuvõte

Praeguseks on olemas programm (autorid Maret Kullasaar ja Evely Nurmsalu), mis võib küll täita nii A kui ka B rolli lihtsas suhtlussituatsioonis, kus A eesmärgiks on, et B otsustaks teha D, kuid nii kasutaja kui ka arvuti saavad valida (kirjutatud) lauseid ainult etteantud nimestikest, st arvuti ei analüüsi ega sünteesi teksti ega kõnet. Seega pole meie mudelis praegu realiseeritud keeleprotsessor ega keeleteadmiste baas TB_K.

Siiani oleme modelleerinud üksnes nõustumise (argumenteerimise) protsessi ja kavandanud mudeli praktilist rakendamist nt suhtlustreeningul, kus arvuti (konversatsiooniagent) saab inimesega suheldes seada teatava distsipliini argumentide ja/või vastuargumentide valiku järjekorrale, millest meie arvates võiks olla kasu argumenteerimisoskuse arendamisel või parandamisel. Kas see tegelikult paika peab, on seni küll veel kontrollimata.

Oma mudeli edasiarendamisel näeme võimaliku rakendusena kõnedialoogsüsteemi. Maailmas on mitmeid eeskujusid selliste süsteemide praktiliseks kasutamiseks näiteks info hankimisel. Selleks tuleb uurida ja modelleerida infohankimisdialoogide struktuuri. Maailma kogemus näitab, et alustada tasub dialoogikorpuse koostamisest, mis ongi meie lähimaks eesmärgiks.

Kirjandus

Ahrenberg, L., Dahlback, N., Jonsson, A. 1995. Coding schemes for studies of natural language dialogue. – Proceedings of the AAAI 1995 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, March. 8–13.

Alexandersson, J., Maier, E., Reithinger, N. 1995. A robust and efficient three-layered dialogue component for a speech-to-speech translation system. – Proceedings of the Seventh European Meeting of the ACL. 188–193.

Allen, J. 1994. Natural Language Understanding. The Benjamins/ Cummings.

Cohen, P 1996. Discourse and Dialogue: Dialogue Modelling. – Survey on the State of Art in Human Language Technology. Toim. P. Cohen. Oregon Graduate Institute.

Dybkjær, L. 2000. Preface. – From Spoken Dialogue to Full Natural Interactive Dialogue – Theory, Empirical Analysis and Evaluation. LREC 2000 Workshop Proceedings. Toim L. Dybkjaer. 1–2.

Feldman, S., Yu, E. 1999. Intelligent *agents*: a primer. – Searcher, Oct 99, Vol 7, 42. Item Number: 2383375.

Fikes, R., Nilsson, N. 1971. STRIPS: a new approach to the application of theorem proving to problem solving. – AI Journal 2, 189–208.

Giahin, E. 1996. Spoken Language Dialogue. – Survey on the State of Art in Human Language Technology. Toim P. Cohen. Oregon Graduate Institute.

Koit, M. 1996. Implementing a dialogue model on the computer. – Estonian in the Changing World. Toim H. Õim. Tartu. 99–114.

Koit, M., Õim, H. 1993. A formal model of communicative strategy. – Proceedings of SCAI-93. 226–231.

Koit, M., Õim, H. 1994. Mõtlemine ja selle mõjutamine tavakujutluses. Kommunikatiivsed strateegiad. – Akadeemia 1, 215–238.

Koit, M., Õim, H. 2000. Developing a model of natural dialogue. – From Spoken Dialogue to Full Natural Interactive Dialogue – Theory, Empirical Analysis and Evaluation. LREC2000 Workshop Proceedings. Toim L. Dybkjær. 18–21.

Lochbaum, K. E. 1998. A collaborative planning model of intentional structure. – Computational Linguistics 24 (4), 525–572.

Nagata, M., Morimoto, T. 1993. An experimental statistical dialogue model to predict the speech act type of the next utterance. – Proceedings of the International Symposium on Spoken Dialogue. 83–86.

Newell, A., Simon, H. 1963. GPS, A program that simulates human thought. – Computers and Thought. Toim E. Feigenbaum, J. Feldman. McGraw-Hill. 279–293.

Stendström, A.-B. 1994. An Introduction to Spoken Interaction. London: Longman.

Sutton, S., Hansen, B., Lander, T., Novick, D. G., Cole, R. 1995. Evaluating the effectiveness of dialogue for an automated spoken questionnaire. – Proceedings of the AAAI 1995 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, March. 156–161.

Walker, M., Whittaker, S. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. – Proceedings of the 28th Meeting of the ACL. 70–78.

Webber, B. 2000. Computational perspectives on discourse and dialogue. – The Handbook of Discourse Analysis. Toim D. Schiffrin, D. Tannen, H. Hamilton. Blackwell Publishers.

Õim, H. 1996. Naïve theories and communicative competense: reasoning in communication. – Estonian in the Changing World. Toim H. Õim. Tartu. 211–231.

Eesti keele tekst–kõne süntees: grafeem–foneem teisendus ja prosoodia modelleerimine*

Meelis Mihkla
Eesti Keele Instituut
Einar Meister, Arvo Eek
Küberneetika Instituut

1. Sissejuhatus

Eesti keele tekst–kõne süntees on Eesti Keele Instituudi, Küberneetika Instituudi ja OÜ Filosofti ühisprojekt. Projekti põhieesmärgiks on luua kvaliteetne kõnesüntesaator, mis teisendaks eestikeelse ortograafilise teksti loomuliku kõlaga ortoeepiliseks kõneks. Kõne akustiliste üksustena kasutame sünteesil difoone.

Et saada tekst–kõne difoonsünteesil arusaadavat, loomuliku kõlaga kõnet, on vaja lahendada viis põhilist ülesannet:

- teksti lingvistiline töötlus,
- grafeem–foneem teisendus,
- kõne prosoodia modelleerimine,
- difoonide andmebaasi loomine,
- digitaalne kõnesüntees.

Teksti lingvistiline töötlus on lahendatud eesti keeletehnoloogide abiga (H.-J. Kaalep, T. Vaino, Ü. Viks, I. Hein). Eesti keele difoonide andmebaas sisaldab üle 1700 difooni (Mihkla, Eek, Meister 1998). Signaalitöötluseks difoonide rittaühendamisel kasutame MBROLA süntesaatorit (Dutoit 1997). Käesolevas töös on põhitähelepanu all eesti keele grafeem–foneem teisendus ja kõne prosoodilise struktuuri modelleerimine.

Kõiki keeli iseloomustab neile omane häälikusüsteeem, hääldusreeglid ja prosoodiline struktuur. Ehkki eesti keelt peetakse nende keelte hulka kuuluvaks, kus õigekirjatähestik sobib suures osas ka kirjakeelse hääldus transkriptsiooniks, esineb eesti keeles küllalt palju kirja ja vastava hääldus kokkusobimatusi. Grafeem–

* Käesolev töö on valminud Eesti Teadusfondi ja Eesti Informaatikakeskuse toetusel.

foneem teisendus esindab reeglite kogumit, mis teisendab ortograafilise teksti häälduste kstiks.

Tekst–kõne süsteemi väljundkõne arusaadavus ja loomulikkus on suuresti sõltuv väljundkõnele rakendatavatest meloodia kontuuridest ja rütmimudelitest. Prosoodiageneraator on vastutav nende kahe külje eest. Töös vaatleme eeldusi ja nõudmisi prosoodiageneraatorile ning käsitleme ka raskusi prosoodiamudeli tuletamisel.

2. Grafeem–foneem teisendus

Lingvistilise keeletöötluse tulemusena teisendatakse ortograafiline tekst häälduste kstiks. Ehkki tavaline tekst tundub eestlasele kergesti häälda tav, valmistab hääldus nii mitte-eestlasele kui ka arvutile raskusi, sest eesti ortograafia ei ole täiesti foneetiline (EKG II 1993). Teksti grafeem–foneem teisendus peab sünteesil tagama eesti keele fonoloogiliselt oluliste vastanduste tajutavuse ja kõne loomulikkuse.

2.1. Diakriitikute automaatne lisamine

1. Kirjapildis ei ole II ja III väldet üldjuhul võimalik eristada (nt *Jaamaesine oli rahvast täis. Rong jõudis jaama.*). Me kasutame kolmanda välte tähistamiseks koolonit (Eek, Meister 1998) eristamaks neid teisevältelistest sama kirjapildiga sõnadest (*Jaamaesine oli rahvast täi:s. Ron:g jõu:dis jaa:ma.*).
2. Samuti on kirjas eristamata palataliseeritud konsonandid palataliseerimata konsonantidest (nt *Tulp on kevadine lill. Tulpdiagramm näitas majanduse kasvu.*). Palatalisatsiooni märkimiseks kasutame apostroofi (*Tul'p on kevadine lill.*).
3. Kõnesünteesil on vaja teada liitsõnapiire (seda tähistab +, nt *elus+olend, kiri+male*), silbipiire (seda tähistab *$*, nt *a$na$lüüs, ka$valamatele$gi*) ja sõnarõhke (% märgib sõna pearõhku, nt *%rääkima*; " märgib kaasrõhku, nt *"mate%maatika*).

3. välte, palatalisatsiooni ja liitsõnapiiri automaatne määramine on lahendatud Filosofti spetsialistide Heiki-Jaan Kaalepi ja Tarmo Vaino abiga. Vastavate diakriitikute asend sõnas määratakse sõnastiku alusel. Väikeses vormisõnastikus (Viks 1992) on 3. välte märk paigutatud rõhulise silbi vokaali ette. Kõnesünteesil on oluline, et 3. välte harjahäälik oleks märgistatud. Selleks, et viia sõnastikupõhine vältemärgi asukoht vastavusse kõnesünteesi vajadustega ja

difoonide andmebaasiga, kasutatakse Ü. Viksi ja A. Eegi koostatud kolmanda välte märgi nihutamise reegleid – fonomalle (vt Tabel 1).

Tabel 1. Kolmanda välte märgi nihutamise reeglid fonomallide alusel
V – vokaal, C – konsonant, Q – kptfš, L – lmnr,
– sõna lõpp, $ – silbipiir, : – 3. välte märk

Fonomall ja nihe	Näide
:VV# →2	v:öö → vöö:
:VVQ# →3	l:aat → laat:
:VVC# →2	v:eel → vee:l
:VVss# →3	p:oiss → pois:s
:VVsQ# →2	l:aast → laa:st
:VVQs# →3	l:oots → loot:s
:VVLQ# →4	h:uult → huult:
:VVCC# →2	k:eeld → kee:ld
:VVCCC# →2	p:aavst → paa:vst
:V# →1	j:a → ja:
:VC# →2	k:as → kas:
:VLQ# →3	k:urt → kurt:
:VLh# →3	mon:arh → monarh:
:VCC# →2	k:ast → kas:t
:VLss# →3	m:arss → mars:s
:VLsQ# →3	k:unst → kuns:t
:VLhv# →3	v:urhv → vurh:v
:VLQC# →3	l:onks → lonk:s
:VCCC# →2	t:ekst → tek:st
:VV$Q →4	l:aa$ta → laa$t:a
:VV$ →2	kr:oo$ni → kroo:$ni
:VVs$s →3	p:ois$se → pois:$se
:VVs$Q →3	l:aas$tu → laas:$tu
:VVL$Q →5	k:aar$ti → kaar$t:i
:VVQ$ →3	k:aat$ri → kaat:$ri
:VVC$ →2	k:eel$du → kee:l$du
:VVQs$ →3	r:oots$lane → root:s$lane
:VVsQ$ →3	s:ääst$l:ik → sääs:t$lik:
:VVLQ$ →4	k:aart$lane → kaart:$lane
:VVCC$ →2	j:uurd$lus → juu:rd$lus
:VVCCC$ →2	p:aavst$lus → paa:vst$lus
:VL$Q →4	k:ar$ta → kar$t:a

Fonomall ja nihe	Näide
:VL$h →4	mon:ar$hi → monar$h:i
:VC$ →2	k:al$du → kal:$du
:VLQ$ →3	p:ilt$l:ik → pilt:$lik:
:VLs$ →3	v:els$ker → vels:$ker
:VLh$ →3	v:urh$vi → vurh:$vi
:VCC$ →2	k:aps$lid → kap:s$lid
:VLsQ$ →3	k:orst$na → kors:t$na
:VLQC$ →3	k:ants$ler → kant:s$ler
:VCCC$ →2	:ekst$ra → ek:stra
:VLQCC$ →3	v:intsk$lema → vint:sk$lema
:VCCCC$ →2	g:angst$rid → gan:gst$rid

2.2. Häälikuteisendused

1) c = *ts*, kui c-le järgneb i või e (Cicero → *tsitsero*)
   c = *k*, kui järgneb mingi muu häälik (curriculum → *kur%rikulum*)
2) w = *v* (Wiiralt → *viiralt*)
3) y = *i* (Kelly → *kelli*)
4) x = *ks* (Xenia → *ksenia*)
5) qu = *kv* (aqua → *akva*)
6) zz = ts (pizza → *pitsa*)

2.3. Hääldusteisendused

1. Pika *üü* diftongeerumine rõhutu silbi lühikese vokaali ees (nt *müüa, hüüe, lüüa* hääldame *müija, hüije, lüija*).
2. *i*-lõpulised diftongid tekitavad rõhutu silbi vokaali ees siirdehääliku *j* (nt *laia, heie, nuia* hääldame *laija, heije, nuija*).
3. Sõnaalguline *g, b, d* muudetakse *k, p, t* –ks (nt *garaaž, buss, diivan* hääldatakse vastavalt *karaa:š, pus':s, tii:van*).
4. *z* ja *ž* asendatakse häälduste kstis vastavalt *s* ja *š* –ga (režii → *rešii:*; zoopark → *soo:+park:*).
5. *n* asendatakse klusiilide *g* ja *k* ees *ŋ* –ga (rong → *roŋ:g*; link → *liŋk:*).

3. Prosoodia modelleerimine

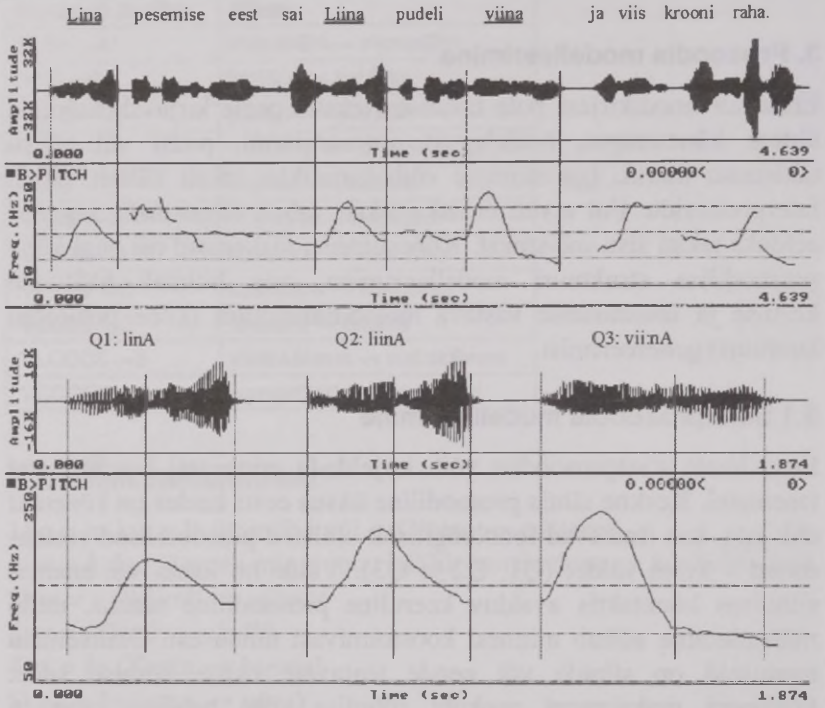
Erinevalt noodikirjast pole tavalises tekstis peale kirjavahemärkide ühtegi kõnetempot, helikõrgust, intonatsiooni, pausi või rõhku tähistavat märki. Iga inimene võib kirjalikku teksti küllalt vabalt interpreteerida. Ent arvuti ei oska teksti vabalt tõlgendada, sest see eeldaks teksti sisu mõistmist. Kõnesünteesi raskemaid osi ongi kõne prosoodilise struktuuri modelleerimine, mis hõlmab häälikute kestuse ja lausetüübile vastava meloodiakontuuri (kõne põhitooni kontuuri) genereerimist.

3.1 Sõnaprosoodia modelleerimine

Eesti keele sõnaprosoodiat võib kirjeldada erinevatel hierarhilistel tasemetel. Keskne tähtis prosoodiline üksus eesti keeles on kõnetakt ehk jalg, kus ilmnevad fonoloogiliselt olulised prosoodilised vastandused – kolm väldet (Q1, Q2 ja Q3). Välde on kahe- või enamasilbilises kõnetaktis avalduv keeruline prosoodiline nähtus, mille määratlemine sõltub mitmest koostoimivast tunnusest. Olulisemaid tunnuseid on silpide või nende teatavate osiste kestuse suhe. Põhitooni maksimumi asukoht rõhulise silbi helilises osas ja akustilise energia jaotus on välte identifitseerimisel teisejärgulise tähtsusega.

Kõnetakt on ka sünteesil põhiüksus sõnaproosoodia modelleerimisel. Joonisel 1 toodud näitelause sisaldab kolme erivältelist sõna: *lina, Liina* ja *vii:na*. Nende sõnade häälikute kestused ja põhitooni kontuurid illustreerivad väldete defineerimist akustiliste parameetrite abil:

1. Kõnetakt on esimeses vältes (Q1), kui takti rõhuline silp lõpeb lühikese vokaaliga ja rõhutu silbi lühike vokaal on foneetiliselt poolpikk või pikk. Põhitooni maksimum on rõhulise silbi helilise osa lõpus ja rõhuta silbi F0 on langev (joonis 1, vasakul).
2. Kõnetakt on teises vältes (Q2), kui rõhuline silp on pikk, st kui ta lõpeb pika vokaaliga, diftongiga või vähemalt ühe konsonandiga. Põhitooni maksimum on rõhulise silbi helilise osa teises pooles (tõusev toon); rõhutus silbis on F0 langev. Rõhutu silbi vokaal on foneetiliselt lühike (joonis 1, keskel).



Joonis 1. Näide kolme välte kohta lauses *Lina pesemise eest sai Liina pudeli viina ja viis krooni raha* (joonise ülaosas on alla kriipsutatud kolm erivältelist sõna). Joonise alaosas on toodud nende sõnade *linA*, Q1, *liinA*, Q2 ja *vii:nA*, Q3 (sõnad esitatud SAMPA transkriptsioonis) kestused ja F0 kontuurid

3. Kõnetakt on kolmandas vältes (Q3), kui rõhuline silp on pikk, st ta lõpeb pika vokaaliga, diftongiga või vähemalt ühe konsonandiga. F0 maksimum on rõhulise silbi helilise osa esimeses pooles (langev toon); F0 langeb jätkuvalt rõhutus silbis. Rõhutu silbi vokaal on eriti lühike ja tugevasti redutseeritud (joonis 1, paremal).

Sõnaprosoodia modelleerimise võib jagada kaheks alamülesandeks:

1) prosoodia andmebaasi koostamine, mis sisaldab andmeid segmentaalsete kestuste kohta ja põhitoonikontuuride mudeleid;
2) reeglite hulga genereerimine segmentaalkestuste ja põhitooni kontuuride juhtimiseks.

Tabel 2. Vokaalide ja konsonantide kestused (meesdiktor)
Paks kiri märgib häälikut, millele esitatud kestus vastab (millisekundites)

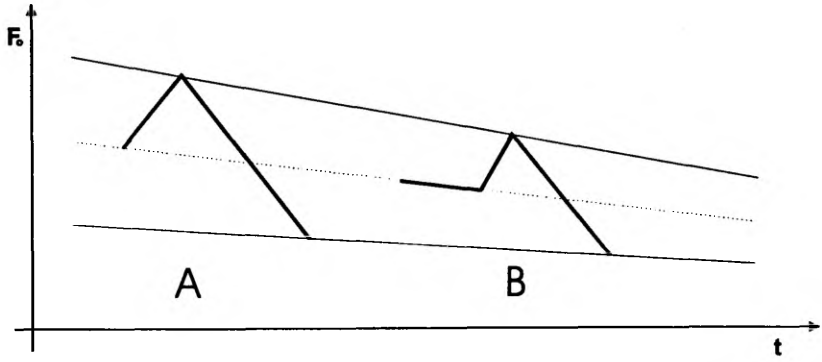
Eriti lühike	Lühike	Poolpikk	Vähendatud pikk	Pikk	Poolteistpikk	Ülipikk
Vokaalide kestused						
60	100	140	175	200	240	300
saade Q3	sade Q1 saade Q2 valede Q1 valet Q1 valeta Q1	sade Q1 valed Q1 valede Q1	saate Q2 loota Q3	saade Q2	saade Q3 loota Q3	maagi Q3
Konsonantide kestused						
	60			120	140–160	180
	pada Q1 saade Q2 saade Q3 valede Q1			kata Q2 saate Q2 loota Q3	loota Q3 valeta Q1	pattu Q3

Prosoodiline andmebaas sisaldab ligi sada kontekstist sõltuvat kestust. Tabel 2 sisaldab kahesilbiliste sõnade vokaalide ja konsonantide kestusi olenevalt takti vältest (Q1, Q2, Q3). Siinjuures peab mainima, et difoonide andmebaas koostati just kahesilbiliste sõnade baasil ja kahesilbiliste sõnade prosoodia on eesti keeles enam uuritud valdkond. Kuid keeles on küllalt palju pikemaid sõnu, mis sisaldavad mitu kahe- või kolmesilbilist kõnetakti. Pikemad sõnad vajavad segmentaalkestuste lühendamist säilitades seejuures väldete kestussuhted.

Tabel 3. Lühikeste vokaalide kestused rõhulises silbis

Järgnev konsonant / Vokaalid	g, b, d	s, z, ž, v, h, m, n, l, r, j	d', n' l', s'
i, ü, u	80	85	90
e, ö, õ	85	90	95
o, ä, a	90	95	100

Tabel 3 esitab näitena rõhulise silbi lühikeste vokaalide kestused, kus erinevused on põhjustatud konsonantide kontekstist ja vokaalide kvaliteedist (nn vokaalide omakestused).



Joonis 2. Sõnade põhitooni kontuurid
A– tavalise eesti keele sõna F0 kontuur
B – võõrsõna F0 kontuur, mil sõnarõhk on teisel silbil

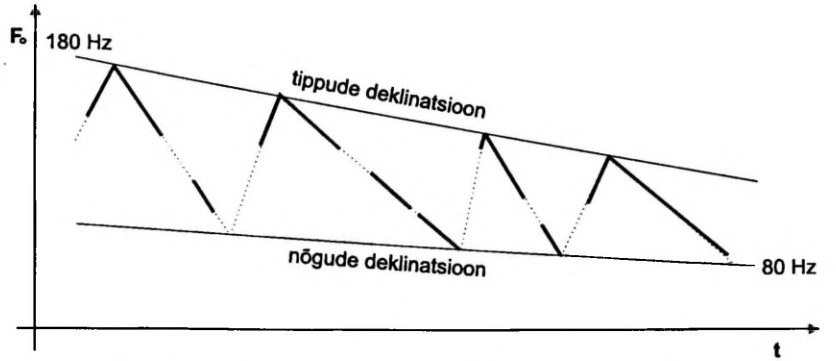
Eesti keeles on sõna esimene silp harilikult rõhuline (st esisilbis asub sõna põhitooni maksimum, mille täpne paiknemine sõltub vältest). Võõrsõnades võib pearõhk langeda esisilbist kaugemalegi (joonis 2).

3.2. Lauseprosoodia modelleerimine

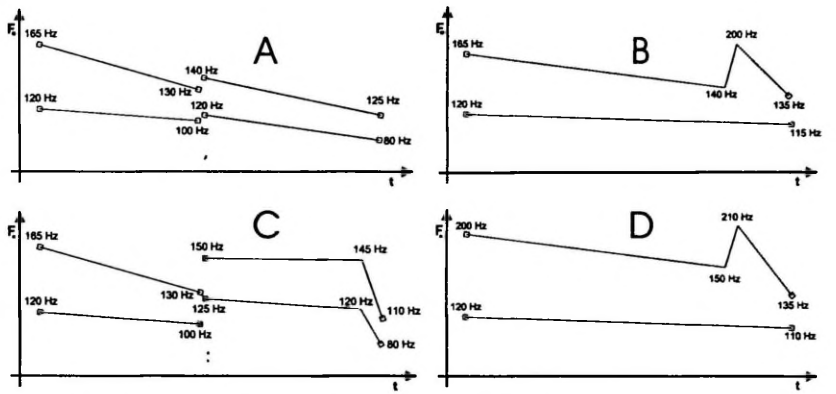
Praeguses sünteesivariandis on intonatsiooni modelleerimise ühikuks lause, mille tüüp määratakse vastavalt lause kirjavahemärkidele. Seega süntesaator loeb teksti lausete kaupa.

Paljudes keeltes on märgatud, et põhitoonikurvidel on kalduvus paikneda keskmiste väärtuste ümber, mis langevad ajas (Vaissiere 1983). Teiste sõnadega, kui arvutada lause keskmine põhitoon, siis lause alguses on F0 väärtused keskmisest kõrgemad ja nad on keskmisest madalamad lause lõpus. Seda põhilist tendentsi märgitakse kui deklinatsiooni. Sõnade põhitoon muutub ajas langevate deklinatsioonijoonte vahel (joonis 3).

Sõltumatult deklinatsioonijoonte arvutamisest saab F0 kõveraid kujutada sihtpunktide jadana, eeldades, et nende punktide vahelised üleminekud täidetakse interpolatsiooni funktsiooniga (nt lineaarse funktsiooni sirgjoontega). Me kasutame nn laia akustilise stiliseerimise meetodit lause meloodiakontuuride modelleerimisel ja nn kitsast stiliseerimise meetodit iga hääliku F0 kontuuri modelleerimisel (Mihkla, Meister, Eek 1999). Iga segmenti iseloomustatakse

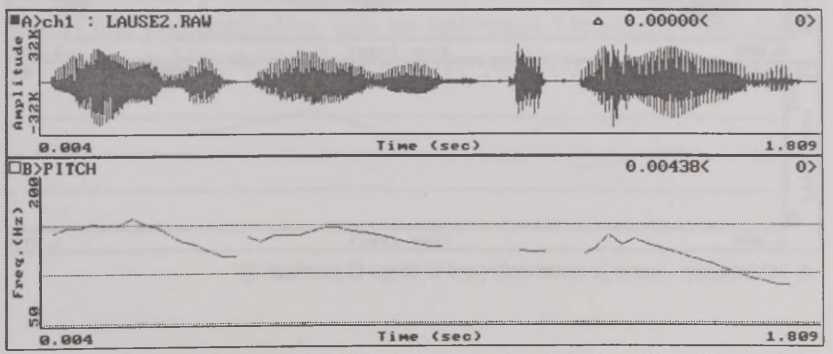


Joonis 3. Põhitooni kõvera F0 akustilise stilisatsiooni näide

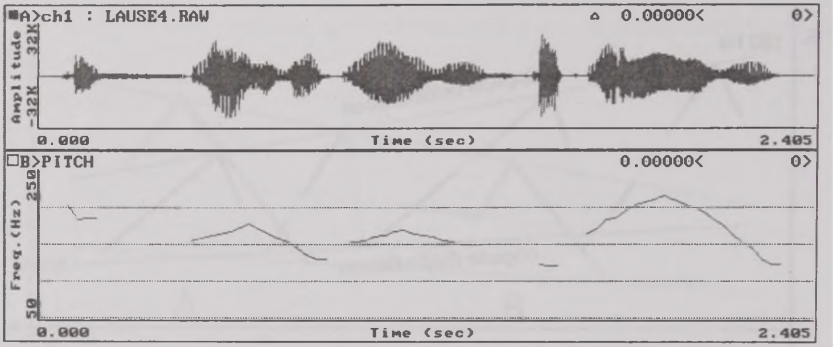


Joonis 4. Erinevat tüüpi lausete deklinatsioonijooned
A – jutustav lause ja neutraalne küsisõnaga küsilause;
B – fokusseeritud küsimus; C – loetelu sisaldav lause; D – hüüdlause

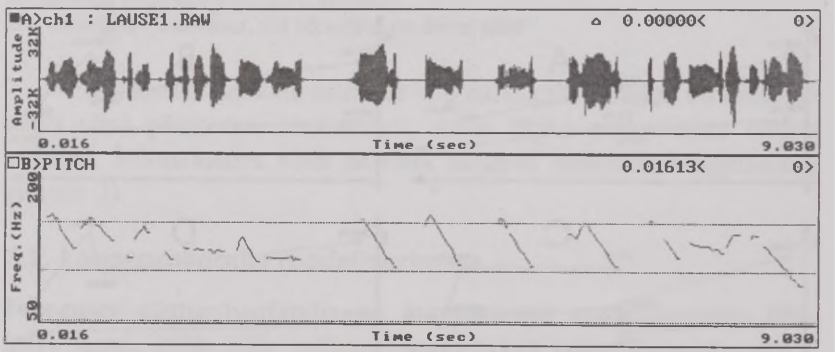
Joonis 5. Sünteesitud lausete F0 kontuuride näiteid



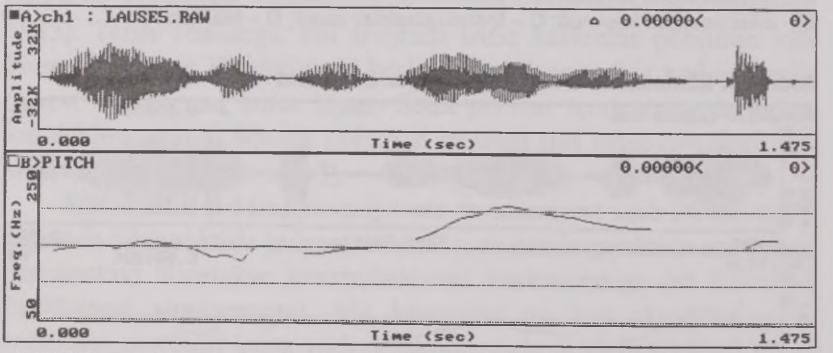
1. Jutustav lause *Jaana joonistab Jaani.* Ja neutraalne küsilause küsisõnaga *Kas Jaana joonistab Jaani?* (vrd. tüübiga A joonisel 4)



2. Fokusseeritud küsimus *Kas Jaana joonistab JAANI?*
(vrd. tüübiga B joonisel 4)



3. Loetelu sisaldav lause *Turul müüakse mitmesuguseid puuvilju: ploome, õunu, pirne, banaane, apelsine ja sidruneid.* (vrd. tüübiga C joonisel 4)



4. Hüüdlause *Jaana ju joonistab!* (vrd tüübiga D joonisel 4)

kolme sihtpunktiga, so. põhitooni väärtus segmendi alguses, keskel ja lõpus.

Laia akustilise stiliseerimise meetodi abil modelleerime erinevat tüüpi lausete deklinatsioonijooni (joonis 4).

Joonisel 5 on kujutatud erinevat tüüpi sünteesitud lausete F0 kontuurid vastavalt joonisel 4 toodud lause mudelitele. Probleemiks on loomuliku ilma küsisõnata küsilause ja hüüdlause prosoodia modelleerimine, sest vastupidiselt neutraalse lause mudelile eeldatakse neil juhtudel lause fookuse leidmist. Kirjeldatud mudel annab suhteliselt häid tulemusi, sest rõhuliste silpide F0 maksimumid ja iga takti kõnerütmi kandvad välte kestussuhted on automaatselt defineeritud.

Lausete jaotamine fraasideks (fraasirõhuga märgistatud sõnade rühmadeks), kus sõnade alluvussuhted on süntaktiliselt defineeritud, parandaks sünteeskõne loomulikkust. Kuid automaatne semantiline-süntaktline analüüs seni veel puudub.

4. Kokkuvõte

Käesolevas artiklis on kirjeldatud eestikeelse tekst–kõne süntesaatoris kasutatavaid grafeem–foneem teisendusi ja prosoodia mudeleid. Sõnaprosoodia tasandil on rahuldavalt realiseeritud kestuste ja põhitooni mudelid ning lauseprosoodia tasandil jutustava, neutraalse küsilause ja loetelu sisaldava lause mudelid. Fokusseeritud küsilause ja hüüdlause prosoodiamudelite, täiustamine eeldab lause fookuse ja fraasirõhkude leidmist, mistõttu sünteesiprotsessi tuleks kaasata ka süntaktiline ja semantiline analüüs.

Grafeem–foneem teisenduse ja prosoodiamudeli algoritmid on realiseeritud dll-programmina (pros.dll) eesti keele tekst–kõne süntesaatori ühe komponendina, mis on priivarana kättesaadav koduleheküljelt www.eki.ee/keeletehnoloogia/projektid/syntees.

Kirjandus

Dutoit, T. 1997. An Introduction to Text-to-Speech Synthesis. Dordrecht: Kluwer.

Eek, A., Meister, E. 1998. Estonian speech in the Babel Multilanguage Database: phonetic-phonological problems revealed in the text corpus. – Proceedings of the Workshop on Speech Database Development for Central and Eastern European Languages. The First International Conference on Language Resources and Evaluation. Granada.

EKG II 1993 = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. Eesti keele grammatika II. Süntaks. Lisa: Kiri. Tallinn: ETA Keele ja Kirjanduse Instituut.

Mihkla, M., Eek, A., Meister, E. 1998. Creation of the diphone database for text-to-speech synthesis. – Proceedings of the Finnic Phonetics Symposium, Pärnu, Estonia, August 11–14, 1998. Linguistica Uralica XXXIV, 3, 334–340.

Mihkla, M., Eek, A., Meister, E. 1999. Text-to-speech synthesis of Estonian. – Proceedings of the 6th European Conference on Speech Communication and Technology, Budapest, Hungary, September 5–10. 1999, vol 5, 2095–2098.

Vaissiere, J. 1983. Prosody: Models and Measurements. Springer-Verlag, Berlin. 53–66.

Viks, Ü. 1992. Väike vormisõnastik. Tallinn.