Haridus- ja Teadusministeerium
Eesti keelenõukogu

EESTI KEELE TEHNOLOOGILISED RESSURSID JA VAHENDID

Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara

Kadri Muischnek
Heili Orav
Heiki-Jaan Kaalep
Haldur Õim

Toimetaja
Urve Talvik

Eesti Keele Sihtasutus
Tallinn 2003

ISBN 9985–79–053–7

SISUKORD

EESSÕNA

Eesti keele arendamise strateegia raames on keeletehnoloogia kui tegevusvaldkonna ülesandeks eesti keeletehnoloogilise toe arendamine sellisele tasemele, et eesti keel oleks võimeline funktsioneerima tänapäeva infoühiskonnas, ühena Euroopa Liidu keeltest. Kui Eesti liitub Euroopa Liiduga, siis on sisulise liitumisprotsessi edukuse üheks eelduseks keeletehnoloogia kõrge tase ja selle arengu tagatus. EL-i programmdokumentide järgi on multikultuuriline ja –keeleline Euroopa võimalik vaid keeletehnoloogiliste vahendite aktiivse kasutuse kaudu. See põhimõtteline seisukoht formuleeriti juba 1990-ndate aastate algul (Danzin 1992). Seni ei ole ükski EL-i liikmesriik loobunud oma riigikeelest mõne rahvusvaheliselt kaalukama liidusisese keele kasuks (kui Iirimaa juhtum kõrvale jätta, aga seal oli juba ajalooliselt ses suhtes eripärane olukord). Eesti keele osas on keeletehnoloogiline tugi seni üpris ebaühtlases seisus, sest pole käivitunud ühtki reguleerivat ja rahastavat programmi, nagu need on tüüpiliselt olemas enamiku Euroopa Liidu keelte puhul.
Keele tehnoloogiline tugi haarab elektroonilisi keeleressursse, keeletöötlustarkvara ja keeletehnoloogilisi rakendussüsteeme. Keeleressursid on elektroonilised andmekogud (tekstid, sõnastikud, andmebaasid), mida kasutatakse keeletarkvara väljatöötamiseks. Keeletarkvara hõlmab keeletöötlemise meetodeid, algoritme ja arvutiprogramme ning on omakorda aluseks keeletehnoloogilistele rakendussüsteemidele alates automaatsetest poolitajatest ja lõpetades kõnetuvastus- ja masintõlkesüsteemidega.

Eesti keele arendamise strateegia siinse lisa eesmärgiks on esitada analüütiline ülevaade keeletehnoloogia nimetatud põhivaldkondadest, nende praegusest seisundist maailmas ja Eestis, ning sellest lähtudes osutada, missuguseid konkreetseid töid tuleb teha, et saavutada strateegia tekstis sõnastatud keeletehnoloogiat puudutavad eesmärgid ja alleesmärgid.

Ülevaade koosneb kolmest osast. Esiteks antakse ülevaade keele arvutikorpustest kui teksti- ja kõneressurssidest, teiseks arvutisõnastikest kui leksikaalsetest ressurssidest ja kolmandaks keeletehnoloogilistest tarkvarasüsteemidest ja nende rakendustest. Igas osas on strateegia koostamise olemusest lähtuvalt kolm alajaotust: 1) ülevaade vastava valdkonna olukorrast maailmas; 2) ülevaade olukorrast Eestis; 3) eelnevast tulenev ülevaade sellest, mida ühes või teises valdkonnas Eestis ja eesti keele osas tuleks teha, et saavutada strateegias fikseeritud eesmärgid.

Arvutikorpuste osa on kirjutanud Kadri Muischnek, arvutisõnastike osa Heili Orav, tarkvarasüsteemide ja nende rakenduste osa Heiki-Jaan Kaalep. Eessõna ja kokkuvõtte on kirjutanud Haldur Õim.

1. KORPUSED

1.1. Korpused ja nende koostamise probleemid

1.1.1. Korpuse mõiste

Keelekorpus on kirjaliku või suulise kõne kogum. Keeleteaduses on sõna *korpus* all enne arvutite kasutuselevõttu tavaliselt mõeldud keeleainese kogumikku, mida kasutatakse uurimistöös materjalina (esineb see siis kartoteegi, lindikogu vms. kujul) vastandina autori enda intuitsioonil põhinevatele üldistustele. Arvutiajastul on korpustena hakatud mõistma peamiselt polüfunktsionaalseid elektroonilisel kujul olevaid tekstikogusid, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest. *Tekst* ei tähenda siin mitte ainult kirjalikku keelt, korpusesse talletatakse ka suulist kõnet. Lühidalt: korpus on loomuliku keele tekstide kogu, mis on koostatud iseloomustamaks keele hetkeseisu või muutumist.

Rangema definitsiooni järgi on tekstikorpus arvutisse viidud tekstide kogum, mis on valitud kindlate kriteeriumite alusel ja esineb ühtses elektroonilises formaadis. Vabama tõlgenduse kohaselt võib korpuseks nimetada ka lihtsalt mingit kogumit tekste elektroonilisel kujul, ehkki tavaliselt säilib ka siin ühtse elektroonilise formaadi nõue. Sel juhul ei ole tekste valitud kindlaid põhimõtteid või eesmärke silmas pidades, vaid neid on kogutud selleks, et kasutaja võiks talletatud tekstide hulgast teha valikuid vastavalt oma vajadustele. Sellist tekstikogumit võib nimetada ka tekstoteegiks või tekstiarhiiviks ja selle tekstid on talletatud seepärast, et igaüks neist on väärtuslik omaette, samas kui korpuse tekstid on väärtuslikud just üheskoos.

Kui esimesed elektroonilised tekstikorpused, nt. Browni tekstikorpus (http://www.hit.uib.no/icame/brown/bcm.html) ja Lancaster-

Oslo/Bergeni tekstikorpus (http://www.hit.uib.no/icame/lob/lob-dir.htm) koosnesid väga läbimõeldud tekstivalikust, aga nende maht oli vaid 1 miljon sõna, siis tänapäeva mitmesaja miljoni sõna suuruste korpuste tekstivalik on märksa juhuslikum. Selle muutuse põhjused peituvad eelkõige arvutitehnika arengus – Browni ja LOBi korpuste koostamise ajal oligi miljon sõna see ülim piir, mis veel arvutimällu mahtus ja mida arvuti töödelda suutis. Tänapäeval mälu probleem praktiliselt puudub, seetõttu saab 'igaks juhuks' tallele panna kõike, mida vähegi jõutakse koguda ja töödelda.

Erinevates allikates kasutataksegi vahel ka termineid *tekstikogu (text collection)* või *tekstiarhiiv (text archive)* tähistamaks sellist elektrooniliste tekstide kogumit, mille valik on olnud juhuslikum. Üldiselt ei toetugi tänapäevaste väga suurte korpuste tekstivalik enam nii täpsetele kriteeriumitele kui paarkümmend aastat tagasi koostatud korpuste omad. Siinkohal on otstarbekas märkida, et arvutilingvistika ja keeletehnoloogia alases kirjanduses nimetatakse sageli korpuseks igasugust tekstide kogu.

Nagu näha, on terminikasutus siin üsna kõikuv. Soovitusi terminite kasutamise kohta võib lugeda nt. internetiaadressil http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html, kus *Expert Advisory Group on Language Annotation Standards (EAGLES)* jagab soovitusi keelekorpuste tüpoloogia ja sellega seotud terminoloogia alal.

Korpusi kasutatakse nii keele uurimisel, sõnaraamatute koostamisel kui ka automaatsete keeletöötlusvahendite väljatöötamisel – seega nii lingvistikas, leksikograafias kui ka arvutilingvistikas ja keeletehnoloogias. Kirjaliku keele korpusi kasutatakse arvutilingvistikas ja keeletehnoloogias näiteks õigekirjakontrollijate (nii ka eesti keele puhul) ja grammatikakontrollijate väljatöötamisel. Korpuste baasil saab luua mitmeid keeletehnoloogia mooduleid nagu formaalseid grammatikaid ja leksikone. Paralleelkorpusi vajatakse tõlkemälu (korduvate üksuste ja nende vastete) koostamiseks ja muude tõlkija abivahendite loomiseks, samuti kontrastiivses lingvistikas.

Võõrkeeleõpetajad kasutavad õpitava keele korpusi selleks, et aidata õpilastel eristada sõna keskseid ja põhilisi tähendusi/kasutusi stilistiliselt markeeritutest.

1.1.2. Korpusi mujal maailmas ja korpuslingvistika üldised arengutendentsid

Esimesed elektroonilised tekstikorpused – juba mainitud Browni korpus USA-s ja Lancaster – Oslo/Bergeni korpus Inglismaal – koostati 60ndatel aastatel. Nende mõlema maht oli üks miljon sõna ja see miljon oli jaotatud tekstiklassidesse, mille osakaal korpuses vastas selle tekstiliigi osakaalule kõigist ameerika või briti inglise keeles kirjutatud tekstidest sellel perioodil. Korpused kajastasid kirjalikku keelekasutust ja sisaldasid kumbki 500 2000-sõnalist tekstikatket. Ligi kakskümmend aastat olid need korpused standardiks, nende korpuste koostamispõhimõtted on olnud eeskujuks paljude teiste korpuste koostamisel, muu hulgas ka tänapäeva eesti kirjakeele baaskorpuse (80ndate korpuse) jaoks. Arvutite mälumahu suurenemine 1990ndate algul võimaldas ka keelekorpuste mahu kiiret kasvu. 1991 alustati Inglismaal kahe suure projektiga – British National Corpus (http://www.hcu.ox.ac.uk/BNC/) ja Bank of English (http://titania.cobuild.collins.co.uk/boe_info.html). Esimene neist (100 miljonit sõna) on representatiivne korpus, so. temasse on tekstid valitud proportsionaalselt nende tekstiklasside esinemisega briti kultuuris. Teine on nn. monitorkorpus, so. sellesse lisatakse tekste pidevalt juurde. Jaanuaris 2002 oli korpuse maht 450 miljonit sõna. BNC-d kasutatakse rohkem mitmekesistel lingvistilistel eesmärkidel, BoE on loodud eelkõige leksikograafide vajadusi silmas pidades. Mõlema korpuse tekstikatked on oluliselt pikemad kui Browni või Lancaster – Oslo/Bergeni korpuste 2000 sõna, nimelt 45 000 ja 70 000 sõna, lühemad tekstid on sisse võetud tervikutena.

Maailma suurim korpus (või korpuste kogum) on praeguse seisuga Mannheimi Saksa Keele Instituudis koostatav *Korpora geschriebener Gegenwartssprache* (http://www.ids- mannheim.de/kt/projekte/korpora/), mis sisaldab umbes 2 miljardit sõna. Parema ülevaate saamiseks asjade seisust maailmas vaatame ka mõningaid väiksemate keelte korpusi.

Sprakbanken (http://spraakbanken.gu.se/) on Göteborgi ülikooli hallatav rootsi keele korpus, mis koosneb umbes 75 miljonist sõnast peamiselt tänapäeva keelematerjalist (ajalehed, ilukirjandus,

bürokraatia keel), aga sisaldab ka keeleajaloolisi tekste. Praegu käib korpuse morfoloogilise märgendamise projekt. Korpus on interneti kaudu vabalt kasutatav.

Suomen kielen tekstipankki on 1990. aastate soome üldkeele korpus, mis koosneb ajakirjandus-, ilukirjandus-, teadus- ja bürokraatiatekstidest, kusjuures ajalehekeel on suures enamuses. Sisaldab ca 180 miljonit sõna. On kasutatav üle interneti *Kielipankki* (http://www.csc.fi/kielipankki/) vahendusel, kuid kasutamiseks on vaja eelnevalt sõlmida leping.

Leedus, Kaunase Vytautas Magnuse ülikooli korpuslingvistika keskuses (http://donelaitis.vdu.lt/indexa.html) on koostatud sajast miljonist sõnast koosnev leedu keele korpus, millest üle poole tekstidest on ajakirjandusest (kuid mitte ainult ajalehtedest vaid ka mitmesugustest ajakirjadest), lisaks ka ilukirjandust (11%), filosoofiaalast kirjandust (3%), Seimi (parlamendi) stenogramme (2%) jpm. Korpus on interneti kaudu vabalt kasutatav.

FIDA, sloveeni keele korpus (http://www.fida.net/eng/) koosneb samuti sajast miljonist sõnast. Korpust saab piiratult kasutada interneti kaudu, tervikkorpuse kasutamine on tasuline.

1.1.3. Korpuste märgendamine

Korpusest on kasu ainult siis, kui saame sealt suhteliselt lihtsalt kätte meile vajaliku info. Aga selleks, et seda vajalikku infot kätte saada, peab sageli alustama info lisamisest korpusesse. Seega: kui soovitakse, et korpus ei jääks ainult elektrooniliste tekstide arhiiviks, tuleb tekstidele lisada info nende ülesehituse kohta (peatükid, pealkirjad, lõigud, laused jne.), samuti andmed morfoloogilise ja süntaktilise analüüsi tulemuste kohta jne. Seda korpusesse info lisamist nimetataksegi korpuse märgendamiseks. Iga teksti juurde tuleb lisada ka bibliograafilised andmed, selle teksti kuulumine mingisse tekstiklassi, teksti pikkus jne. Tavaliselt lisatakse tekstifaili algusesse nn. päis (ingl. k. *header*), kus need asjad dokumenteeritakse. Märgendada saab täisautomaatselt, käsitsi (ka arvuti interaktiivse abiga) või neid kahte meetodit kombineerides. Näiteks saab eestikeelset teksti

morfoloogiliselt märgendada (so. teostada morfoloogilist analüüsi ja ühestamist) peaaegu täisautomaatselt, kuid süntaktiline märgendamine toimub käsitsi, kasutades visualiseerivat ja antud kontekstis võimalikke variante väljapakkuvat tarkvara.

Igasugune märgendamine algab teksti jagamisest lõikudeks ja lauseteks (või mingiteks lausesarnasteks üksusteks). Seda tehakse tavaliselt täisautomaatselt, ka eesti keele jaoks on olemas üsna häid tulemusi andev lausestaja.

Siinkohal tuleb tingimata rõhutada korpuse korduva kasutamise võimalust. Näiteks morfoloogiliste kategooriate, nagu teistegi lingvistiliste kategooriate identifitseerimist ei saa käsitleda omaette, isoleerituna. Teiste sõnadega, ei saa analüüsida ainult osa meile vajalikke sõnu tekstis ja teistele, meile momendil mittevajalikele mitte oma aega ja jõudu raisata. Kui kord kas nt. morfoloogilise või süntaktilise analüüsi tulemused on korpuse tekstidele lisatud, siis selle töö tulemusena saadud märgendatud korpus on juba palju väärtuslikum ressurss kui märgendamata korpus, seda saavad kasutada paljud väga erinevatel eesmärkidel. Korduva kasutamise võimalus on väga oluline, sest märgendamine on tavaliselt väga kallis ja aeganõudev tegevus.

Heal juhul saab korralikult märgendatud korpust kasutada paljudel erinevatel eesmärkidel. Näiteks on morfoloogiline analüüs ka mitmete teiste analüüsiliikide, nt. süntaktilise ja semantilise analüüsi alus, morfoloogiliselt märgendatud korpusest saab otsida sõna kogu paradigmat korraga ja ainult morfoloogiliselt märgendatud korpuse põhjal saab teha sõnade sagedussõnastikke ja muud sõnastatistikat.

Maailmas enim levinud märgendustasemed ongi *morfoloogiline* ja *süntaktiline* märgendamine. Selliselt märgendatud korpused on ka eesti keeles kõige mahukamad: morfoloogiliselt on 2003. aasta lõpu seisuga märgendatud 600 000 sõna ja süntaktiliselt 200 000 sõna. Senine süntaktiline märgendamine on toimunud vastavalt kitsenduste grammatika formalismile (vt. nt. Müürisep 2000). Maailmas on lisaks sellele formalismile enamlevinud fraasipuude ja sõltuvuspuude esitamine. Tuntuim fraasistruktuuri suhtes märgendatud korpus (puude pank) on *Penn Treebank* (http://www.cis.upenn.edu/

~treebank/), suurim saksakeelne sellealane projekt on TIGER (http://www.ims.uni-stuttgart.de/projekte/TIGER/). Viimasel ajal kogub üha enam kuulsust Praha Sõltuvuspuude Pank (*The Prague Dependency Treebank*; http://quest.ms.mff.cuni.cz/pdt/index.html).

Semantiline märgendamine seisneb selles, et igale sõnale lisatakse teave selle sõna kuulumise kohta mingisse semantilisse klassi. Semantiline klass või semantiline väli on teoreetiline konstruktsioon, mis ühendab neid sõnu, mida mingil üldisemal tasemel saab liita ühe mõiste alla. Lihtsamas variandis mõeldakse semantilise märgendamise all sõnatähenduste ühestamist. Eesti keeles on semantiliselt märgendatud 100 000 tekstisõna. Ühestatud on nimisõnade ja verbide tähendused kasutades eesti wordnet'i (http://www.cl.ut/ee/ressursid/teksaurus.html).

1.2. Eesti keele korpused

1.2.1. Eesti keele korpused Tartu Ülikoolis

Tartu Ülikooli korpused on internetis kättesaadavad aadressidel http://www.cl.ut.ee ja http://www.murre.ut.ee

1.2.1.1. Kirjutatud keele korpused

Tänapäeva eesti kirjakeele korpuse tegemist alustati nn. baaskorpusest (80ndate aastate korpusest), mis on analoogiline inglise keele klassikaliste tekstikorpuste – Browni ja Lancaster- Oslo/Bergeni (LOB) korpustega (nende kohta vt p.1.1.1 ja p.1.1.2 käesolev ülevaade). Eesti kirjakeele baaskorpuses on tekstid aastatest 1984 – 1987, kokku 1 miljon sõna. Tekstid on jaotatud kümnesse tekstiklassi (vt. tabel 1), mis esindavad representatiivselt selle aja kirjalikke tekste. Igast valitud tekstist on korpusesse võetud 2000-sõnaline katkend, kui tekst oli lühem (ajaleheartiklid), siis on ta korpusesse võetud tervikuna. Sellest korpusest on välja jäetud tõlked, lastekirjandus, draama ja luule. Korpuse struktuuri ja koostamispõhimõte-

te kohta saab põhjalikumat infot artiklitest Hennoste, Muischnek jt. 1993; Hennoste 1996; Hennoste, Koit jt. 1998; Hennoste, Muischnek 2000; Hennoste, Kaalep jt. 2001.

Valdkond	**Sõnade arv**	**Protsent korpusest**
Ajakirjandus	175 000	17,5
Dokumendid	12 000	1,2
Entsüklopeedilised teosed	20 000	2,0
Esseed ja biograafiad	90 000	9,0
Hobid ja harrastused	75 000	7,5
Ilukirjandus	250 000	25,0
Populaarteadus	150 000	15,0
Propaganda	60 000	6,0
Vaimulikud tekstid	8 000	0,8
Teadus	160 000	16,0

Tabel 1. Tekstiklassid tänapäeva eesti kirjakeele baaskorpuses.

Selle miljonisõnalise korpusega liituvad nn. läbilõikekorpused perioodist 1890–1990, täpsemalt ilukirjandustekstid aastatest 1890–1899 (155 000 sõna), 1900–1910 (64 500 sõna), 1911–1920 (247 000 sõna), 1935–1939 (252 000 sõna), 1945–1954 (66 000 sõna), 1966–1970 (257 000 sõna) ja 1988–1998 (611 000 sõna) ning ajakirjandustekstid aastatest 1890–1899 (193 000 sõna), 1900–1910 (171 500 sõna), 1911–1920 (182 500 sõna), 1935–1939 (117 000 sõna), 1948–1952 (242 400 sõna), 1966–1970 (168 500 sõna) ja 1988–1998 (385 000 sõna).

Seega on baaskorpuses ja läbilõikekorpustes kokku umbes 4 miljonit sõna.

Samal internetileheküljel on väljas ka esimesed tekstid (ca 36 miljonit sõna) loodavast suurest tekstikorpusest, mille maht loodetakse lähiaastatel viia 100 miljoni sõnani, aastaks 2010 aga 200 miljonini. Sellesse korpusesse võetakse terviktekstid, mitte tekstikatked. Nagu eesti kirjakeele varasematesse korpustessegi, ei lülitata siia luulet ega draamat. Korpuse koostamisel ei panda enam niivõrd rõhku representatiivsusele, kuivõrd tekstide (õigemini sõnade)

suurele hulgale, et iga kasutaja saaks siit teha valikuid vastavalt oma vajadustele. See loodav korpus saab olema avatud monitorkorpus, kuhu uusi tekste pidevalt juurde lisatakse. Suhteliselt lihtne on koguda ajalehekeelt, vastavad lepingud on juba sõlmitud enamuse eesti suuremate lehtedega (Postimees, Eesti Päevaleht, Eesti Ekspress, Maaleht) ja korpusese ajalehetekstide maht suureneb 2003. aasta lõpuks 50 miljoni sõnani ning juba sõlmitud lepingute raames võib suureneda veel 100 miljonini. Kõik maailma väga suured korpused koosnevad suuremas osas ajalehekeelest ja seda peetakse üldkeele suhtes üsna neutraalseks keelekasutuseks. Kuid siiski peaks ideaalis vähemalt 10% korpuse mahust (ehk 10 miljonit sõna) moodustama ilukirjandus ja teist sama palju teadustekstid. Ilukirjanduse kogumine on kõige raskem ja seda mitte niivõrd tehnilistel põhjustel, vaid seetõttu, et ilukirjandusteose autoriõiguste omanikult või omanikelt on kõige raskem saada nõusolekut teksti korpusesse lülitamiseks. Teadustekstide autorid loovutavad oma tekste korpusesse meelsasti, siin seisneb põhiprobleem selles, et eesti keeles kirjutatakse vähe täppis- ja tehnikateaduste tekste.

Loodavast korpusest on praegu interneti kaudu kättesaadavad:

- 13 miljonit sõna Riigikogu toimetatud stenogramme,
- ajalehed “Postimees” ja “Eesti Ekspress”, kokku 11,4 miljonit sõna,
- Eesti ja Euroopa Liidu seadusandlikud aktid ca 11 miljonit sõna.

Kõik need korpused on interneti kaudu vabalt kättesaadavad mittekommertsiaalsetel eesmärkidel. Üks teksti korpusesse lülitamise eeltingimus ongi see, et teksti autoriõiguste omanikuga on sõlmitud leping, mis lubab neid tekste selliselt kasutada.

Korpuse koostajad on arvamusel, et ei ole mõtet koostada tekstikorpust nn. sisemiseks kasutamiseks – vaid kümnele-kahekümnele inimesele.

Lehekülje www.cl.ut.ee tänapäeva eesti keele korpust ja läbilõikekorpuseid saab kasutada ühtse kasutajaliidese abil. **Kasutajaliides** on programm, mis abistab kasutajat korpusest info otsimisel, seega siis programm, mis otsib korpusest vajalikud keelekasutuse näited

ja esitab need koos kontekstiga. Nendes eesti keele korpustes on kontekstiks alati lause (selle eelduseks on märgendatud lausepiirid). Praegu on nendest korpustest võimalik otsida sõnu ainult sellisel kujul, nagu nad on tekstis, so. sõnavormidena. Selleks, et saaks tekstist otsida mingi sõna kõiki vorme, peavad tekstid olema morfoloogiliselt märgendatud. Sellisele tekstile saaks esitada ka päringut grammatilise kategooria kohta, nt. "anna laused, mis sisaldavad nimisõnu komitatiivis".

1.2.1.2. Paralleelkorpused

Paralleelkorpus on korpus, mis sisaldab mingit teksti originaalkeeles ja selle tõlget teise keelde või ka tõlkeid teistesse keeltesse. Paralleelkorpuste abil uuritakse tõlkeprotsessi ja selle kaudu luuakse mitmesuguseid tõlkija abivahendeid. Keeletehnoloogias on masintõlkesüsteemide väljaarendamiseks samuti vajalikud väga suured paralleelkorpused.

Selleks, et paralleelkorpusest ka kasu oleks, tuleb teda eelnevalt rohkem töödelda kui tavalist ükskeelset tekstikorpust. Paralleelkorpused paralleelistatakse, so. mingid omavahel vastavuses olevad eri keelte üksused (laused, osalaused, fraasid, sõnad) märgendatakse üksteisele viitavalt.

Ilmselt tuntuim paralleelkorpus maailmas on Kanada *Hansard*, mis koosneb Kanada parlamendidebattidest, mida avaldatakse riigi kahes ametlikus keeles – inglise ja prantsuse keeles.

Väiksemate keelte hulgast võib jällegi näiteks tuua sloveeni keele, millel on 1 miljoni sõnaline inglise-sloveeni paralleelkorpus (http://nl.ijs.si/elan/).

Selliseid paralleelkorpusi, kus üheks keeleks on eesti keel, on selgelt liiga vähe. Europrojekti Multext-East (http://nl.ijs.si/ME/) tulemusena on eesti keel üheks keeleks paralleelkorpuses, mis sisaldab George Orwelli romaani "1984" kaheksas keeles. Selle korpuse maht on 75 000 sõna ja paralleelistatud on ta lause tasandil. Eestikeelne osa sellest on kättesaadav http://www.cl.ut.ee/ee/1984/, paralleelkorpust levitatakse CD-l.

Eesti keelt sisaldab ka paralleelkorpus SCLOMB, milles on kogutud Läänemere-äärsete keelte tekste (ilukirjandust) ja nende tekstide tõlkeid teistesse Läänemere-äärsetesse keeltesse. Korpus on koostatud Soomes Turu ülikoolis, seda levitatakse CD-l, interneti kaudu kättesaadav ei ole.

Lisaks paralleelkorpustele kasutatakse veel ka *võrreldavaid korpusi (comparable corpora),* mis sisaldavad tekste eri keeltes (või ka keelevariantides) samal teemal. Tuntuim võrreldav korpus on ICE – International Corpus of English (http://www.ucl.ac.uk/english-usage/ice), mis sisaldab 20 inglise keele varianti kas ingliskeelsetest maadest nagu Suurbritannia või USA või siis riikidest, kus inglise keel on teiseks ametlikuks keeleks või kõrghariduse keeleks nagu näiteks Indias. Eesmärgiks on inglise keele erinevate regionaalsete variantide kontrastiivne uurimine.

1.2.1.3. Vana kirjakeele korpus

Eesti vana kirjakeele korpust on TÜ vana kirjakeele uurimisrühm (http://www.murre.ut.ee/vakkur/) koostanud alates 1995. aastast. Korpuses on praegu eesti kirjakeele vanimad tekstid aastani 1600, Georg Mülleri jutlused, Turu käsikiri ja Rossihniuse kirikukäsiraamat. Eelpoolnimetatud teoste maht kokku on u. 250 000 sõna. Arvutisse on viidud ka kõik Stahli teosed, 2003. aasta oktoobri seisuga on need veel lemmatiseerimata, kuid lähiajal saab ka neid interneti kaudu kasutada. Plaanis on koostada vana kirjakeele valikkorpus, mis kataks ajavahemikku 1535 – 1850. Töö selle kallal algab 2003. aasta lõpupoole.

Korpuse põhjal valminud sõnastike ja teiste publikatsioonide nimekirja leiab aadressilt http://www.murre.ut.ee/vakkur/Yllitised/yllitised.htm .

Korpusele saab esitada päringuid interneti kaudu. Kuna kõik korpuse tekstid on morfoloogiliselt märgendatud, nimelt varustatud infoga lemma ja sõnaliigi kohta, saab päringuid esitada lemma kohta.

1.2.1.4. Eesti murrete korpus

Eesti murrete korpus (http://www.murre.ut.ee/murded_index.html) on projekt, mis sai alguse 1998. aastal. Murdekorpus valmib Tartu Ülikooli eesti keele õppetooli ja Eesti Keele Instituudi koostöös. Tööd on seni tehtud Eesti Teadusfondi (grant nr 4192 "Eesti murrete elektrooniline andmebaas II" juht J. Viikberg, EKI) ja teadusteema „Eesti kirjakeele arenemine ja varieerumine ning selle murdetaust" (teema juht TÜ professor M. Erelt) toel.

Korpus koosneb kõige vanematest lindistustest, mis on piisava kvaliteediga. Kasutatud tekstid on lindistatud valdavalt 1960–1970ndatel, need on litereeritud enamasti Eesti Keele Instituudis ning kontrollitud ja sisestatud Tartu Ülikoolis.

Tekste on kogutud kõigilt murdealadelt, igalt murdealadelt on valitud 2–3 murrakut, kust on pärit põhiosa tekstidest.

2003. aasta lõpuks valmiv 0,5 miljoni tekstisõna suurune kogum moodustab murdekorpuse põhiosa, millele hiljem on võimalik lisada tekste kas teistest kihelkondadest või hilisemast ajaperioodist (nn. siirdkeelt). Korpuse põhiosa saab kasutada nii foneetilises kui lihtsustatud transkriptsioonis ning see peab saama tervikuna morfoloogiliselt märgendatud.

Murdekorpust on üritatud arendada ühtlaselt, s.t. kõigi murrete ja murrakute tekstid on olnud töös paralleelselt. 2003. aasta oktoobris on murdekorpuses kokku 455 900 litereeritud tekstisõna kõigist Eesti murretest (Liina Lindströmi andmed):

Murre	**Murrak**	**Sõnu**	**Märgendatud**
Idamurre	Torma	10080	
Idamurre	Kodavere	10419	
Keskmurre	Viru-Jaagupi	10458	
Keskmurre	Väike-Maarja	5848	
Keskmurre	Keila	7338	
Keskmurre	Juuru	7990	
Keskmurre	Jüri	341	
Keskmurre	Pilistvere	13204	
Alutaguse	Lüganuse	22656	4108

Alutaguse	Jõhvi	13894	
Rannamurre	Jõelähtme	22975	2974
Rannamurre	Kuusalu	20930	
Saarte murre	Käina	27249	
Saarte murre	Kihelkonna	23161	
Saarte murre	Mustjala	857	
Saarte murre	Pühalepa	14097	
Saarte murre	Kihnu	23400	
Läänemurre	Varbla	18790	
Läänemurre	Mihkli	16225	
Läänemurre	Häädemeeste	5610	
Mulgi murre	Karksi	13793	
Mulgi murre	Tarvastu	4358	
Mulgi murre	Halliste	5752	
Tartu murre	Rõngu	19994	3730
Tartu murre	Otepää	13986	
Tartu murre	Nõo	2262	
Tartu murre	Võnnu	20564	
Tartu murre	Kambja	10876	
Võru murre	Hargla	17711	13488
Võru murre	Urvaste	5091	
Võru murre	Põlva	2199	1633
Võru murre	Räpina	4370	3970
Võru murre	Vastseliina	17203	11261
Võru murre	Setu	42219	19249
Kokku		**455900**	**60413**

Ebaühtlus eri murrete tekstisõnade hulgas on tingitud muredelindistuste ja litereeringute olemasolust ning kättesaadavusest. Valminud on tekstide morfoloogilist märgendamist abistav programm. Oktoobriks 2003 oli morfoloogiliselt märgendatud 60 413 tekstisõna. Põhirõhk on seni olnud lõunaeesti (peamiselt Võru) tekstide märgendamisel, teiste murdealade tekstide morfoloogiline märgendamine on kavas lähiaastatel. Juba märgendatud tekstidest päringute sooritamiseks on valmimas internetipõhine otsinguprogramm. Eesti murrete korpuse põhjalikuma iseloomustuse leiab nt. artiklites Lindström jt. 2001 ja Lindström 2001.

1.2.1.5. Suulise kõne korpus

Klassikaline keeleuurimine on olnud eelkõige kirjalike tekstide keele uurimine. Alates 20. sajandi 60ndatest aastatest on uuritud ka suulist kõnet. Eesti keele suulise kõne korpust (http://sys130.psych.ut.ee/~linds/) on tehtud TÜ üldkeeleteaduse õppetooli juures alates 1996. aastast. See on planeeritud avatud korpusena, st. ta piirsuurust ei ole määratud. Siia on mõeldud koguda erinevat tüüpi suulist kõnet, nii argisuhtluse kui avaliku suhtluse keelekasutust, nii spontaanset kui ettevalmistatud kõnet, nii monolooge kui dialooge. Lindistused on litereeritud Jeffersoni transkriptsiooni kasutades, mis on ette nähtud eelkõige vestluse, keelelise suhtluse uurimiseks.

Septembris 2003 oli suulise kõne uurimisrühma tekstikogus Andriela Rääbise andmetel:

- 440 linti, mis on umbes 2000 tundi lindistatud vestlusi;
- 832 transkribeeritud teksti, mis on umbes 606 500 tekstisõna.

Sellest silmast-silma vestlusi 364 teksti (306900 sõna), neist 117 argivestlust ja 247 institutsionaalset vestlust ja telefonivestlusi 389 teksti (195000 sõna), neist 126 argivestlust ja 263 institutsionaalset vestlust. Raadio- ja TV-saateid on litereeritud 79 teksti (104 600 sõna).

Suulise kõne litereeringutest on morfoloogiliselt märgendatud 100 000 sõna: 60 000 sõna argivestlusi ja 40 000 ametlikke vestlusi.

Väike osa suulise kõne korpusest on kättesaadav ka interneti vahendusel, kuid enamikku korpusest saab teadustööks kasutada ainult lepingu sõlmimisel. Lepingu vormi leiab suulise kõne korpuse internetileheküljelt. Linte ning litereeringuid saab teadustööks Andriela Rääbise käest (andriela@ut.ee). Erinevalt kirjaliku keele korpusest on suulise kõne korpuse lindistuste või tekstide puhul probleemiks kõneluses osalejate privaatsuse säilitamine, lisaks veel nende privaatsuse kaitse, kellest nendes vestlustes räägitakse. Nende probleemide tõttu pole ka tulevikus plaanis kogu korpust interneti kaudu kasutatavaks muuta.

Suulise kõne korpuse kohta vt. ka Hennoste, Lindström jt. 2000; Hennoste 2000 ja Hennoste, Lindström jt. 2001.

1.2.1.6. Dialoogikorpus

Dialoogikorpust saab kasutada nii inimestevahelise suhtluse uurimiseks kui ka inimesega loomulikus keeles suhtleva dialoogsüsteemi (arvutiprogrammi) arendamiseks.

Eesti dialoogikorpus sisaldab seisuga oktoober 2003:

1) eesti suulise kõne korpusest 277 dialoogi, kokku 100 000 tekstisõna, neist 150 telefonikõnet ja 105 silmast silma vestlust.
   50 000 tekstisõna ulatuses on märgendatud dialoogiaktid (tervitus, palve, avatud küsimus jms.).
2) „Võlur Ozi" meetodil (kus inimese ja arvuti dialoogis simuleerib arvutit teine inimene) on kogutud 21 (kirjalikku) (reisiinfo)dialoogi, kokku 3400 sõna. Kõigis on märgendatud dialoogiaktid.

Kogutud materjal on konfidentsiaalne ega ole avalikult kasutatav.

Dialoogikorpuse suulisi dialooge haldab TÜ suulise kõne uurimisrühm ning simuleeritud dialooge TÜ arvutiteaduse instituut. Korpus on praegu alles koostamisjärgus ning tal puudub kasutajaliides; selle loomine on aga edaspidi plaanis.

Dialoogikorpuse loomise probleemidest on juttu artiklites Hennoste jt. 2002, Koit 2002a, 2002b, 2003.

1.2.2. Eesti Keele Instituudi korpus

Eesti Keele Instituudi korpus on hetkel suurim eesti tänapäeva kirjakeele korpus, millele saab interneti kaudu päringuid esitada. Ta on kättesaadav internetileheküljel http://www.eki.ee/corpus/. Korpuse maht on umbes 13 miljonit sõna, sellest enamuse moodustavad ajalehetekstid. Tekstid ei ole mingil moel märgendatud, sobides eelkõige leksikaalse materjali otsinguks.

Korpus jaguneb järgmisteks tekstiklassideks:

1) ajalehed umbes 10 miljonit sõna;
2) ajakirjad (ajakirjad, populaarteaduslikud väljaanded, seadusetekstid jms.) 0,3 miljonit sõna;

3) ilukirjandus umbes 2 miljonit sõna;
4) (pseudo)kõnekeel ca 50 000 sõna: 14 osa teksti seriaalist „Dallas“.

1.3. Eesti keeletehnoloogia vajadused korpuste osas

Kõikides keeletehnoloogia valdkondades ei jõuta ilmselt eesti keele tarbeks luua oma tehnoloogiaid. Sellel poleks ka mõtet, sest maailmas on juba palju keelest sõltumatuid programme. Selleks, et neid eesti keele jaoks kohandada, läheb vaja väga suuri korpusi.

1.3.1. Kirjutatud keele korpus

Esmane vajadus (eelkõige keeletehnoloogide ja leksikograafide, aga kindlasti ka lingvistide jaoks) on tõeliselt **suur korpus**, mis koosneks minimaalselt 100 miljonist sõnast ja sisaldaks võimalikult erinevaid tekstiklasse – lingvistika vajadusteks vähemalt ilukirjandus, ajakirjandus ja nn. akadeemiline keelekasutus; keeletehnoloogia vajadusteks seda tüüpi tekstid, mille jaoks keeletehnoloogilisi tooteid looma hakatakse (nt. töötav s.o. toortõlget väljaandev masintõlkesüsteem saab tänapäeval olla ainult väga valdkonnaspetsiifiline). Eesti keele arendamise strateegia näeb ette kasvatada korpuse maht aastaks 2010 vähemalt 200 miljoni sõnani.

1.3.1.1. Paralleelkorpused

Keeletehnoloogia, aga ka kontrastiivse lingvistika vajaduste jaoks on tingimata vajalik **suur paralleelkorpus**, mis sisaldaks vähemalt paarkümmend miljonit sõna. Kuid tõsise masintõlke-alase töö jaoks läheb vaja 100 miljoni sõna suurust paralleelkorpust.

1.3.1.2. Spetsiaalkorpused

Vastavalt keelestrateegiale kuuluvad eesti keeletehnoloogia lähiaja ülesannete hulka muu hulgas süntaktilise süvastruktuuri analüsaator ja loomulikku keelt kasutavad liidesed, ka kõnesisendiga liidesed. Nende jaoks läheb vaja vastavalt süntaktilise süvastruktuuri suhtes märgendatud korpust (vähemalt 100 000 sõna) ja dialoogiaktide suhtes märgendatud korpust. Nagu näeme, tekitavad keeletehnoloogia ees seisvad uued ülesanded ka vajaduse uute spetsiaalkorpuste järele.

1.3.2. Korpuste märgendamine

Selleks, et juba olemasolevaid ressursse paremini ära kasutada, on vaja kõik interneti kaudu kasutatavad tänapäeva kirjakeele korpused morfoloogiliselt **märgendada**. Kuna on olemas 600 000 sõna morfoloogiliselt perfektselt märgendatud tekste, mida saab kasutada statistikapõhise morfoloogilise ühestaja treenimiseks ja testimiseks, siis on alust arvata, et ülejäänud korpuste morfoloogiline märgendamine saab toimuda juba automaatselt. Vanema keelekasutuse ja suulise kõne automaatne morfoloogiline märgendamine ei ole ilma suure lisatööta tõenäoline.

Ka võimalikult suure hulga tekstide süntaktiline (nii kitsenduste grammatika (Müürisep 1998a, 1998b, 1999) kui ka mingi fraasistruktuuri esitava formalismi järgi) ja semantiline märgendamine on vajalik nii vastavate keeletöötlusprogrammide arendamiseks kui ka lingvistilise uurimistöö tarbeks.

1.3.3. Ühtne kasutajaliides

Mõistlik oleks kõik eesti tänapäeva kirjakeelt sisaldavad korpused ühendada ühe **kasutajaliidese** alla. Kusjuures see kasutajaliides peaks võimaldama (lisaks konkordantside so. sõna kontekstis otsimisele):

- teha korpuste põhjal mitmesugust statistikat (nt. mitu korda küsitud sõna, järjend või lingvistiline üksus esineb kogu korpuses / teaduskeeles / juriidilistes tekstides jne.);
- teha päringuid sõna algvormi järgi (so. kogu materjal peab olema morfoloogiliselt märgendatud);
- lisaks üksiksõnale peaks saama otsida ka mitmesõnalisi ühendeid (ka neid, mille liikmed tekstis kõrvuti ei asetse);
- saada infot kollokatsioonide (sõnade koosesinemuste) kohta;
- korpuse alaosa (minimaalselt 50 000 sõna) peaks olema ka süntaktiliselt märgendatud.

1.3.4. Suulise kõne korpus ja kõneandmebaasid

Suulise kõne korpus oma praegusel kujul on kasutatav vaid lingvistilistel eesmärkidel. Keeletehnoloogias kasutamiseks läheks aga vaja sellist suulise kõne korpust, kus kõne ja transkriptsioon on omavahel paralleelistatud.

Et anda eesti keelele teiste keeltega võrdseid võimalusi suhtlemiseks infotehnoloogilises keskkonnas, on vajalik käivitada arendustööd eestikeelse kõne tuvastustehnoloogia väljatöötamiseks. Kõnetuvastusprogrammi väljatöötamiseks tuleb luua eestikeelne kõne väljavõtteid sisaldav andmebaas (2000 kõnelejat), mis on ilmtingimata vajalik nii uuringute kui arendustööde teostamiseks kui ka tuvastussüsteemi treenimiseks ning testimiseks. Vastav projekt (http://www.phon.ioc.ee/base) käivitati TTÜ Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris 2002. aastal (Meister jt. 2002, Meister 2003).

2. ARVUTILEKSIKONID

2.1. Sissejuhatus

Siinse osa eesmärk on anda ülevaade arvutileksikonide kui keeleressursside arengust maailmas; sellest, missuguseid olulisemaid elektroonilisi leksikone Eestis olemas on, milliseid on juurde vaja ning missuguseid neist oleme Eestis oma praegusi teadmisi ja kogemusi arvestades suutelised realiseerima.

Loomuliku kõne ja keele uurimisega tegelejad on jõudnud arusaamisele, et töökindlate ja tõhusate keeletehnoloogiliste toodete loomine sõltub otsustavalt sellest, kui kättesaadavad on suured adekvaatsed keeleressursid, nimelt elektroonilised sõnastikud, terminoloogiabaasid, teksti- ja kõnekorpused ning formaalsed grammatikad. Iga rakendus, mis kasutab sõnu, vajab ka arvutisõnastikke ning tihti on just sõnastik süsteemi kesksek osaks. Elektroonilised sõnastikud erinevad traditsioonilistest (paber)sõnastikest. Samuti ei vaja kõik arvutirakendused ühesuguseid sõnastikke: õigekirjakorrektori jaoks on tarvis mahukat sõnastikku, mis ülesehituselt võib olla üsna lihtne, masintõlkesüsteem nõuab aga eeskätt just detailirikast, paindlikku ja keeruka struktuuriga sõnastikku. Samuti vajavad eri süsteemid eri liiki infot – masintõlge ja teksti mõistmise süsteemid kasutavad sõnastikes esitatud infot (vasted, seletused), kõnetöötlus vajab aga sõnastikes olevat häädusinfot.

On oluline teha terminoloogilist vahet:

- arvutis loetav sõnastik (*machine readable dictionary*, MRD) – s.o. traditsioonilise sõnastiku arvutiversioon,
- arvutileksikon (*computer lexicon*, CL) – s.o. arvutisõnastik, mis pole koostatud (ainult) paberversiooni väljalaskeks.

See eristus ei tähenda, et arvutis loetav sõnastik ei saaks olla väärtuslikuks materjaliks arvutileksikoni koostamisel.

On kaht tüüpi arvutisõnastikke: ühed leksikonid on mõeldud inimeste jaoks; teised on suuremate rakendussüsteemide osad. Inimeste sõnastikele on tehtud kasutajaliidesed, mille abil kasutaja saab infot sõnade kohta. Rakendussüsteemidesse kuuluvad sõnastikud töötavad muude süsteemide osana. Leksikaalse info süstemaatilise kirjeldusega tuleb tegeleda kummalgi juhul.

2.2. Ülevaade arvutileksikonide arengust

Arvutite kasutamine sõnavara uurimises ja sõnastike koostamises sai laiemalt võimalikuks pärast seda, kui 50. – 60. aastatel oli hakatud arvutisse viima suuremaid tekstikogusid, korpusi. 1964. aastal valmis *Browni* korpus, 1978 *Lancaster-Oslo-Bergen* korpus (vt. lähemalt käesoleva ülevaate osa 1. Korpused). Samuti hakati koostama ka teiste keelte korpusi. Korpused said leksikograafide ressurssideks näit. *Svensk Ordbok* (1987); *Collins Cobuild Dictionary Bank of English* põhjal (1987). Korpuste põhjal koostati konkordantse ja indekseid, nt. KWIC (*Key Word In Context*) indekseid, kus sõnade esinemused on toodud koos tekstilõiguga, kus vastav sõna esineb. Seda materjali kasutati leksikoloogilises analüüsis.

Üheks esimeseks alaks, kus piisavalt suured tekstimaterjalid osutusid väga efektiivseteks vahenditeks, oli sõnade esinemussageduste uurimine, tulemuste põhjal sagedussõnastike koostamine, aga ka sageduste põhjal mitmesuguste oluliste järelduste tegemine sõna positsiooni kohta keele leksikaalses süsteemis.

Teine liin arvutite tulekuks leksikoloogiasse ja leksikograafiasse oli otseselt sõnastike koostamine arvutil ja olemasolevate sõnastike viimine arvutikujule. Üks esimesi tuntud sõnastikke, mis koostati arvutil ja mida levitati ka elektroonilises versioonis, oli *Longman Dictionary of Contemporary English* ehk *LDOCE* 1978. a. Järgnesid inglise keele sõnaraamatud/andmebaasid: *Oxford English Dictionary*, *Webster's 7th, Collins English Dictionary, The Penguin English Dictionary, Oxford Advanced Learner's Dictionary of Current English, The New Shorter Oxford English Dictionary on Historical Principles* ja mitmed kirjastuse Collins kakskeelsed sõnastikud.

Arvutileksikonide loomise käigus kerkis 80. – 90. aastate vahetusel küsimus: kas jätkata nii nagu seni, koostades iga uue rakenduse jaoks oma leksikon, või üritada välja töötada põhimõtted leksikonide loomiseks ja leksikaalse materjali esitamiseks, mis võimaldaksid vältida dubleerimisi ja asjatuid kulutusi. Tänapäeval on teine lähenemine üldaktsepteeritud. Selle taustaks on keeletehnoloogias juurdunud **keeleressursside korduvkasutatavuse** nõue. Keeleressursside korduvkasutatavuse nõudest tulenevalt on info esitusviisi standardiseerimine üks olulisemaid ülesandeid. Üldistatud märgenduskeele rahvusvaheliseks standardiks kinnitati 1986. aastal *Standard Generalized Markup Language,* lühendatult SGML (ISO Standard 8879). Üldistatud märgenduse sünonüüm ongi SGML. SGML-i üheks edasiarenduseks on rahvusvahelise uurimisprojekti *Text Encoding Initiative* (TEI) poolt välja töötatud kodeerimisskeemid ehk märgendusmudelid. Need on valmis mudelid ehk märgendikomplektid (ingl. k. *tag sets)* paljude erinevate tekstitüüpide jaoks. Erinevus SGML-ga võrreldes: TEI loodi keeleressursside jaoks. Märgendatud on korpusetekste, trükitud sõnaraamatuid jm.

Oxford English Dictionary (täpsemalt *New OED* ehk *OED2*) on ehk kuulsaim 1980-ndate arvutileksikograafiline töö. *OED* jaoks kujundati välja spetsiaalne andmemudel, teksti märgendamiseks kasutati tollal uudset SGML-i, info otsimiseks töötati välja oma päringusüsteem. Kokkuvõttes on *OED* struktureeritud kui SGML-märgenduses tekstiandmebaas. Hiigelprojekt teostati Kanadas, Waterloo ülikoolis, kus arvutisse sisestati *OED* 13 köidet ja 4 lisaköidet (kokku 500 000 sõnaartiklit, sh. 1,8 miljonit näitetsitaati). Mammuttöö tulemusena ilmus 1989. aastal *OED* 2. trükk.

Longman Dictionary of Current English on samuti üles ehitatud SGML-ile.

Arvuti kasutamise ühe suuna – keelandmete kogumise ja sorteerimise – ammendas 1980-ndate alguse COBUlLD-i sõnastikuprojekt. Projekti raames loodi 20 miljonist sõnast koosnev inglise tekstikorpus, sõnu sorditi kõikvõimalikesse konkordantsidesse ning saadud alusmaterjalile toetudes koostati COBUlLD-i sõnaraamat.

1980-ndatel leiti, et on vaja andmebaasi vormi, mis oleks kasulik automaatselt taksonoomiate, seletuste jms. tegemiseks. Arvutis

loetavaid sõnastikke hakati kasutama erinevate semantiliste hierarhiate ehitamiseks. Võtmesõnaks sai *leksikaalne andmebaas.*

2.2.1. Leksikaalsed andmebaasid

Leksikaalse andmebaasi (ingl.k. *lexical database,* LDB, ka *machine-tractable dictionary,* MTD) all mõistetakse arvutileksikoni, kus nii selles sisalduvad andmed kui ka selle struktuur on esitatud täiesti eksplitsiitselt ning tänu sellele on võimalik koostada paindlikult liigendatud päringuid.

Ka **semantilised andmebaasid** on tegelikult leksikaalsete andmebaaside alaliik selles mõttes, et tegeldakse tüüpiliselt sõnadega. Kuid semantilistes andmebaasides on põhirõhk sõnade tähenduste ja eriti sõnadevaheliste semantiliste seoste kajastamisel.

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

- **Tesaurus** on tavatähenduses mõistelise sõnaraamatu liik, kus sõnavaraüksused ei ole organiseeritud mitte alfabeetiliselt vaid sisuseoseid pidi. Tesaurusele on omane hierarhiline struktuur ja alluvussuhted mõistete vahel.
- **Arvutitesaurus** tähendab andmebaasi elektroonilisel kandjal, kus sisaldub info keeleüksuste ja nendevaheliste sisuseoste kohta. Andmebaasiga liitub kasutajaliides, mille abil tesauruse kasutaja saab kätte selle osa informatsioonist, mis on talle vajalik. Kasutajaliideselt eeldatakse ka liikumisvõimalust tesauruse ühelt sõlmelt teisele. Arvutitesaurus võib olla personaalselt kasutatav (CD-l) või võrgu kaudu kättesaadav.

Tuntumaid leksikaal-semantilisi andmebaase on WordNet (WN) (Fellbaum 1998; http://www.cogsci.princeton.edu/~wn/), mille loomist alustati 1980ndate aastate keskel. WordNet oli algselt mõeldud realiseerima (ja kontrollima) teatud ideid inimese mentaalse leksikoni ehituse kohta. Eeldati, et sisend leksikoni on mitte sõnavormide, vaid tähenduste kaudu. Seetõttu on WordNet organiseeritud mitte sõnade järgi nagu tüüpiline sõnastik või leksikaalne and-

mebaas, vaid tähenduste järgi, kusjuures tähendused on esitatud seda tähendust väljendavate sünonüümide loendiga e. sünohulkadega (ingl. k. *synsets*).

Näide: 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…*

EuroWordNet (EWN) (http://www.illc.uva.nl/EuroWordNet/) 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.
EuroWordNeti peamine erinevus WordNetist ongi tema mitmekeelsus. Kõik projektis osalejad lõid WordNeti põhimõttelisele ülesehitusele toetudes omakeelse wordneti, kus keeltevahelise indeksi (*interlingual index,* ILI) kaudu on võimalik leida sama mõistet väljendavad sünonüümihulgad teistes keeltes.

Tähendused (so. sünohulgad) on asetatud üksteisega leksikaal-semantilistesse seostesse, ühtekokku ligi 60 erinevat suhtetüüpi. Olulisemad semantilised seosed on:

- hüponüümia/hüperonüümia (nt. *inimene-elusolend*),
- troponüümia (vastab verbide puhul hüponüümiaseosele, nt. *kõndima-marssima*),
- meronüümia e. osa –tervikuseos (nt. *auto-rool*),
- antonüümia (*pikk-lühike*),
- järgnevusseos (seob eelkõige verbide tähendusi, nt. *norskama-magama*) jt.

Nende seoste kaudu moodustavad sünohulgad hierarhiaid. Hierarhiad on eriti levinud nimisõnade tähendustes, vähem verbidel, veel vähem omadussõnadel.

Sarnaste leksikaalsemantiliste ressursside olemasolu paljudes keeltes võib viia mitmete heade tulemusteni. Automaatsed tõlkesõnastikud on ainult üks neist. Samuti on wordnet-tüüpi tesaurus kasulik intelligentsetele info-otsisüsteemidele, mis on võimelised otsima mõisteid või tähendusi mitmetes erinevates keeltes.

2.2.2. Leksikaalsed teadmusbaasid

Leksikaalsete andmebaaside kõrval on üha enam hakatud rääkima ka **leksikaalsetest teadmusbaasidest.** Üks peamisi erinevusi leksikaalsete teadmusbaaside ja leksikaalsete andmebaaside vahel on esimeste võime esile tuua üldistusi ja tuletada järeldusi. Leksikaalne andmebaas võimaldab lihtsalt esitada andmeid sõnahaaval ning teeb võimalikuks nende andmete otsimise. Näiteks on inimese jaoks tavaline, et sõnad nagu *klaas, kruus, kann* võivad tähistada mitte ainult teatud nõusid, vaid ka vedeliku kogust, mis neisse mahub. See on kogu vastava semantilise sõnaklassi üldine omadus ja vastavalt peaks selline üldistus – selle võimalikkus – ka arvutileksikonis kajastuma. Leksikaalne andmebaas seda ei võimalda.

Nagu öeldud, on arvutileksikonid arenenud teoreetilise lingvistika leksikoni- ja üldisemalt semantikakontseptsioonide mõjul. See seos tiheneb kahtlemata veelgi tulevikus, eriti näiteks leksikaalsete teadmusbaaside loomisel, kus on tarvis teoreetilises semantikas välja töötatud üldistus- ja järeldusmehhanisme. Tutvustame seda seost teoreetilisest lingvistikast välja kasvanud **freimisemantika** (Saluveer, Õim 1985; Õim, Saluveer 2002) näite varal.

Freimi mõiste toodi keeleteadusesse teatud kindlas ideoloogilises kontekstis: freimides nähti eelkõige vahendeid, mille abil puhtkeelelisi teadmisi saaks siduda relevantsete mittekeeleliste argiteadmistega. Konkreetsemalt on freimide kasutamist seni seostatud leksikoniga, sõnade tähenduste kirjeldamisega. Seetõttu on lõviosa teoreetilistest diskussioonidest ja ka üksiknäidete käsitlustest seotud freimide käsitlemisega leksikaalse semantika vahendina – freimid kui leksikaalse semantika uus kontseptuaalne vahend, freimid kui vahend, mille abil sõnade tähenduste kirjeldustesse saab sisse tuua relevantseid argiteadmisi, mis seostuvad sõna poolt tähistatava situatsiooniga (Fillmore 1977).

FrameNet on freimisemantikal baseeruv projekt, mille käivitas Charles Fillmore California Ülikoolist Berkeleys 1997. a. (http://www.icsi.berkeley.edu/~framenet).

Projekti eesmärk on luua leksikaalne andmebaas viie tuhande ingliskeelse sõna jaoks, mis kataksid erinevaid semantilisi valdkondi.

Iga kirje peaks näitama sõna omadusi ja kasutust, mida tõestatakse sõna esinemisega 100 miljoni sõnalises korpuses. Kirjeldatavad sõnad on kogutud semantiliselt seotud valdkondadest. Need valdkonnad projektis on: muutus, tunnetus, suhtlus, emotsioonid, tervis, liikumine, taju, sotsiaalne keskkond, ruum, eluetapid, aeg, äritegevus. Need üldised freimid omakorda jagunevad allfreimideks.

Kasutades kindlaksmääratud märgendeid, kontrollivad uurijad süstemaatiliselt iga sõna kasutust erinevates korpustes ja leiavad näiteid andmebaasi jaoks. Leksikograafid valmistavad lõpuks ette sisendi lõpliku kuju, mis sisaldab lemmat, viidet tema freimile, esinemisvõimaluste loendit ja iga võimaluse kohta illustreerivat näitelauset. Kogu selle töö lihtsustamiseks on loonud programmeerijad tarkvara.

Allikmaterjalina kasutatakse projektis WordNeti andmebaasi (20 000 sõna koos omavaheliste semantiliste suhetega), arvutisõnastikke (trükitud sõnastike elektroonilised versioonid), COMLEX andmebaasi (38 000 sõna markeeritud süntaktiliste tunnustega) (http://cs.nyu.edu/cs/projects/proteus/comlex/).

Freimisemantika ideede teise realisatsioonina võib kirjeldada projekti **DELIS** ("*Descriptive Lexical Specifications and tools for corpus-based lexicon building*"). See on mahult märksa piiratum, kuid see-eest rohkem sügavuti semantikasse suunatud. DELIS on Euroopa Liidu projekt (Heid, Krüger 1996; http://www. hltcentral. org/projects/detail.php?acronym=DELIS). Projektis DELIS on tehtud väike paralleelne verbileksikon 5 Euroopa keele jaoks (inglise, prantsuse, itaalia, taani ja hollandi keel). DELIS kasutab leksikaalsete üksuste omaduste kirjeldamiseks HPSG süntaksi ja freimisemantika meetodeid. Praegune DELIS'e leksikon kirjeldab eelkõige suhtlus- ja tajuverbide freime. DELIS'e projektis väljatöötatud esitusviisi kasutab ka eelpool nimetatud FrameNet projekt.

2.2.3. Kokkuvõte

Arvutileksikograafia areng on kulgenud arvutisse sisestatud sõnastikutekstidelt leksikaalsete andmebaasideni. Tulevikku jääb leksikaalsete teadmusbaaside koostamine.

Arvutis loetavaid sõnastikke on maailmas palju. Neid on nii akadeemiliseks kasutuseks kui ka ärilise kasusaamise eesmärgil välja antud. Viimaseid on müüdud ka CD-ROM-idel. Näiteks *CD-OED* sai valmis 1992, soome *CD-Perussanakirja* 1997, pakutakse *LDOCE*, *Webster'i* jt. CD-sid.

Kõige kriitilisem punkt mitte ainult leksikonide koostamiseks, vaid ka loomuliku keele töötluse arendajate, süsteemiinseneride jne. jaoks on **standardiseerimise vajadus.** Selleks on vaja:

- et info oleks hästi struktureeritud ning täpselt märgendatud mingis kodeerimissüsteemis (SGML; TEI),
- et süsteemil oleks hea juurdepääs võimalusega andmeid lisada ning luua erinevaid viiteid (linke).

Euroopa Liidu keeletehnoloogia programmi raames on üks ulatuslikumaid projekte EAGLES (*Expert Advisory Group on Language Engineering Standards*, vt. ka p.1.1.1. käesolev ülevaade), mille eesmärgiks on välja töötada standardid mitte ainult leksikonide, vaid ka korpuste jt. keeleressursside jaoks.

Teine kriitiline punkt on juriidiline **autoriõiguse** küsimus. Üsna paljud uurijad pole saanud kirjastajatelt luba teha uurimusi nende sõnastike või andmebaaside põhjal, sest kirjastused on huvitatud kasu saamisest. See aga pole valitsev. Leidub mitmeid kirjastusi, kes on teinud lepingud enda materjalide avalikuks kasutamiseks.

Autoriõiguse küsimus aga kahtlemata jääb. Nende küsimustega tegelemiseks on näiteks ACL (*Association for Computational Linguistics*) teinud ettepaneku asutada CLR (*Consortium for Lexical Research*), mille eesmärk on aidata teostada uurimusi loomuliku keele sõnastikest ja leksikonidest ning samas teha leksikaalsed andmekogud ning töövahendid kättesaadavaks laiemale üldsusele. Erinevate keelte puhul võib situatsioon olla muidugi väga erinev.

Arvuti ja sõnastike probleeme arutatakse arvutilingvistika konverentsidel (COLING) ja leksikograafiakonverentsidel (EURALEX, COMPLEX).

2.3. Mis on Eestis juba olemas

Eestis on sõnastikud suhteliselt hästi esindatud. Olemas on nii üldkeelealased, kakskeelsed kui ka erialasõnastikud, neist paljudest on olemas ka arvutiversioon.

2.3.1. Tänapäeva eesti üldkeelealased sõnastikud

Enamasti on Eestis levinud paljude pabersõnastike elektrooniline variant. Hulk elektroonilisi sõnastikke on kasutatavad ka interneti kaudu. 1998. aastal avanes KeeleWeb (http://ee.www.ee), mis koondab erinevaid keeleasju. KeeleWeb on mittetulunduslik arendusprojekt, mille tegevuse otseseks eesmärgiks on eesti keele sõnastike, tekstikorpuste ja keeletarkvara kättesaadavaks tegemine internetis ning nendega seotud teenuste arendamine. KeeleWeb püüab jõudumööda kaasa aidata eestikeelse tarkvarakeskkonna tekkimisele. Projekt sai alguse Avatud Eesti Fondi abirahast 1997, nüüdseks on KeeleWeb ennast sidunud Eesti Keele Sihtasutusega.

Järgnevalt nimekiri teadaolevatest arvutis olevatest eesti üldkeele sõnastikest (vt. ka Langemets 2002):

- Õigekeelsussõnaraamat (1976). 114 000 märksõna. Sõnaraamatu materjal viidi arvutisse 1980. a. Otsimootor internetis KeeleWebi kodulehel (http://ee.www.ee/QS/). On 2 eri keerukusega otsimisvõimalust: 1) lihtpäring (annab transkriptsioonimärkidega või liitsõnapiiriga kuju, sõnaliigi, tüübinumbri ja tüüpsõna) ning 2) komplekspäring (otsib paljude üksikute väljade järgi, võimaldades ka nende kombineerimist).
- A. Õim. Antonüümisõnastik (1995). Ligi 2000 eesti üldkeele märksõna koos oma vastandiga, mis koos moodustavad antonüümipaari. Otsimootor internetis KeeleWebi kodulehel (http://ee.www.ee/Anton/).
- A. Õim. Fraseoloogiasõnaraamat (1993). 6500 püsiväljendit. Püsiväljend võib olla nii mitmest sõnast koosnev ühend kui ka liitsõna. Otsimootor internetis KeeleWebi kodulehel (http://ee.www.ee/Fras/).

- A. Õim. Sünonüümisõnastik (1991). 10 000 sünonüümirida. Otsimootor internetis KeeleWebi kodulehel (http://ee.www.ee/Synon/).
- A. Saareste. Eesti keele mõistelise sõnaraamatu indeks (Uppsala, 1979). 132 000 sõna. Indeks ei ole sõnaraamat ise, vaid raamatu kasutamiseks hiljem koostatud hädavajalik loend (selle elektrooniline ja interneti-variant tehti Eesti Keele Instituudis (EKI)). Otsimootor internetis EKI kodulehel (http://www.eki.ee/dict/saareste/). Internetis saab vaadata iga sõna juurest ka "Väiksesse murdesõnastikku" ja sealt omakorda levikukaarti. Samuti saab kohe liikuda A. Rauna "Eesti keele etümoloogilise teatmiku" info juurde.
- A. Raun. Eesti keele etümoloogiline teatmik (Brampton–Tartu, 1982). 5800 sõna (sh. ca 500 viidet). Otsimootor internetis EKI kodulehel http://www.eki.ee/dict/raun/). EKI-s tehtud interneti-variandile on lisatud ka sõnade loendid märgendite (nt. deskriptiivne, mütoloogiline) ja isikute järgi, kes midagi on väitnud või loonud (nt. J. Aavik, P. Ariste, F. J. Wiedemann).
- P. Päll. Maailma kohanimed (1999). 4200 artiklit, kokku ligi 16 000 nimevarianti. Osaliselt (nt. maailma maade ja pealinnade nimed ning ISO maatähised) tekst internetis (http://www.eki.ee/knab/mkn_ind.htm). Interneti-lisadena pakub autor näha sõnastiku parandusi ja täiendusi ning süstemaatilise registri ehk kohanimed riikide kaupa.
- Ü. Viks, J. Sang. Riimisõnastik. 600 000 sõnavormi, ca 175 000 riimipesa. EKI-s tekstiandmebaasina, mille põhjal on plaanitud kirjastada CD-versioon.
- Ü. Viks. Väike vormisõnastik (1992). 46 000 märksõna (sh. viited). See on eesti keele esimene spetsiaalne morfoloogiasõnastik. Andmebaas koos morfoloogilise analüüsi ja sünteesiga on saadaval EKI-s.
- Eesti keele sõnaraamat (1999). 50 000 sõnaartiklit. EKI-s struktuurimärgenditega lihttekst. Plaanitud on (koos paranduste ja täiendustega) teha ka interneti-versioon.
- Eesti kirjakeele seletussõnaraamat I–VII. (1988–; tegemisel: viimane vihik ilmub 2007). Seni ilmunud 102 000 sõnaartiklit.

EKI-s struktuurimärgenditega lihttekst. Andmebaasiks on teinud selle Leho Paldre, mida kasutatakse Tartu Ülikooli arvutilingvistika uurimisrühmas.

- Filosofti tesaurus (1997). Sisaldab 60 000 sõna, 13 000 sünonüümihulka. Koostatud A. Õimu "Sünonüümisõnastiku" ja "Antonüümisõnastiku" põhjal. Otsimootor internetis KeeleWebi kodulehel (http://ee.www.ee/Tesa/).
- TEA Võõrsõnastik (1999). Üle 30 000 sõna. Müügil nii CD-ROMil kui ka otsimootor internetis KeeleWebi kodulehel (http://ee.www.ee/VS/).
- M. Loog. Esimene eesti slängisõnaraamat (1991). 7500 sõna. Otsimootor internetis KeeleWebi kodulehel (http://ee.www.ee/dict/slang/). Internetis saab (märk)sõnu otsida tähejärjendi või teemanumbri järgi.
- Väike murdesõnastik I–II (1982–1989). 73 000 märksõna. Otsimootor internetis EKI kodulehel (http://www.eki.ee/dict/vms/). Otsiprogramm töötab kahes eri režiimis – teksti ja leviku alusel –, mõlemal juhul saadakse vastuseks vastavad sõnad koos levikualaga.
- Hargla murraku morfoloogiline andmebaas. See on ühe eesti murraku grammatiline andmebaas, mis tulevikus võiks kujuneda üldiseks murdegrammatika baasiks. Otsimootor internetis EKI kodulehel (http://www.eki.ee/dict/hargla/).
- F. J. Wiedemann. Eesti-saksa sõnaraamat (1973). EKI-s sisestatud struktuurimärgenditega lihttekst; otsimootor internetis EKI kodulehel (http://www.eki.ee/dict/wie/).
- Eestis alustati EuroWordNet'i (vt. eespool) projekti raames eesti üldkeele tesauruse e. eesti wordneti (EstWN) koostamist 1997. aastal TÜ arvutilingvistika uurimisrühmas. Sünohulki on EstWN-s hetkel (oktoober 2003) u. 11,5 tuhat (tesaurus täieneb pidevalt) – põhiliselt substantiivi- (66%) ja verbimõisted (27%), kuid vähesel hulgal ka adjektiive (2,6%) ja pärisnimesid (4,4%). Semantilisi seoseid on ühel sünohulgal EstWN-s üle kahe, domineerivad hüpo- ja hüperonüümiasuhted. Eesti keele tesauruse andmebaas eksisteerib EWN andmebaasina (keelest sõltuv moodul) Polarise formaadis TÜ arvutilingvistika uurimisrühmas.

Otsimootor TÜ arvutuslingvistika kodulehel (http://www.cl.ut.ee/ee/ressursid/teksaurus.html).

- H. Kaalep, K. Muischnek. Eesti kirjakeele sagedussõnastik (2002). Sagedussõnastiku aluseks on üks miljon sõna ajakirjandust ja ilukirjandust. Esitatud on sõnad, mis esinesid mõlemas tekstiklassis ja kokku vähemalt viis korda. Sõnade algvormid e. lemmad leiti automaatselt, kasutades eesti keele morfoloogilist analüsaatorit koos statistilise ühestajaga. Olemas elektrooniline versioon TÜ-s, interneti versioon TÜ arvutilingvistika kodulehel (http://www.cl.ut.ee/ee/tulemusi/sagedused.html).
- On ka mitu ilmunud sõnaraamatut, mille kogu elektroonilisus seisneb esialgu nende küljendamises arvuti abil, nagu näiteks suur "Eesti murrete sõnaraamat I–XI" (1994–; tegemisel: ilmunud on 11 vihikut), M. Kallasmaa "Saaremaa kohanimed I–II" (1996–2000), V. Palli "Idamurde sõnastik" (1994).

Elektroonilised sõnastikud on oluliselt toetanud paljude keeletöötluslike ülesannete lahendusi, nagu süntaktiline analüüs, sõna tähenduse valik, kõnesüntees, tekstianalüüs, infokorraldus, fraasianalüüs jpm. Eestis on viljakaim olnud Ü. Viksi (1992) "Väike vormisõnastik", millest lähtub kogu eesti keele automaatne morfoloogiline analüüs, nii EKI morfoloogiline analüüs ja süntees kui ka tarkvarafirma Filosofti poolt pakutavad mitmesugused eesti keele analüüsi vahendid: speller (kontrollib õigekirja), morfoloogiline analüsaator, lemmatiseerija (leiab eesti keele sõnade algvormid), süntesaator (käänab või pöörab) ja poolitaja. Ka semantilist infot – täpsemini: leksikaalseid suhteid sõnade vahel – on üle võetud olemasolevatest elektroonilistest sõnaraamatutest, nt. Filosofti tesaurus on tuge saanud sünonüümi- ja antonüümisõnastikust, eesti wordnet Eesti Kirjakeele Seletussõnaraamatust jm.

2.3.2. Oskussõnastikud ja terminibaasid

Oskussõnastikke on mitmesuguseid. Suuremad on valminud TA asutustes ja vormistavad varem valminud materjali uuel kujul, väiksemad on entusiastide kätetöö. Erialadest on paremini esindatud zooloogia ja bioloogia ning arvutite ja tarkvaraga seonduv.

Järgnevalt antakse nimekiri oskussõnastikest võrgus. Toodud võrguoskussõnastike loend asub aadressil http://ee.www.ee/oskus ja teda täiendatakse uute andmete saamisel. Nimestiku on koostanud Arvi Tavast.

- 187 mõistet toidu, toitumise, toitainete ja enesehoolduse vallast: abivahend kõigile (2000). Koostajad M. Zilmer, U. Kokassaar, T. Vihalemm. (http://www.parnu.ee/raulpage/kokk/187.html),
- Vello Hanson, Arvi Tavast. Arvutikasutaja sõnastik (1999). 5926 kirjet (http://ee.www.ee/AKS/),
- Vello Hanson. Arvutisõnastik (1995). 8438 mõistet. (http://www.ioc.ee/arvutisonastik/),
- Urmas Laansoo, Sulev Savisaar, Ülle Reier, Jaak Palumets. Eestikeelsete taimenimede andmebaas (2002). 10600 nimepaari. (http://www.ut.ee/taimenimed/),
- Eesti õigusterminite sõnastik (2001). 4202 kirjet. Eesti Õigustõlke Keskus (http://www.legaltext.ee/),
- ESTERM – Eesti Õigustõlke Keskuse terminibaas (2002). 41747 kirjet. Eesti Õigustõlke Keskus. (http://www.legaltext.ee/),
- Heikki Vallaste. e-teatmik (2002). 3437 kirjet. (http://www.vallaste.ee/eteatmik)
- Euroopa Liidu eelarve sõnastik (2001). 815 märksõna. Peatoimetaja T.-T. Reinbusch. Eesti Õigustõlke Keskus. (http://www.legaltext.ee/),
- Euroopa Liidu õigusterminite sõnastik (2001). Eesti Õigustõlke Keskus. (http://www.legaltext.ee/),
- EVS-ISO/IEC 2382. Infotehnoloogia. Sõnastik. (standardiprojekti töövariant) (1999). (http://www.imprimaatur.ee/standard/sisukord.htm),
- Tiia Haud. Galaktilise astronoomia ja kosmoloogia sõnastik (2002). Sagedussõnastik. (http://www.aai.ee/~tiia/sonastik),

- Mart Viikmaa. Klassikalise geneetika leksikon (1998). (http://madli.ut.ee/~martv/genolex.html),
- Migratsioonialane oskussõnastik (2000). Kodakondsus- ja Migratsiooniamet, Eesti Õigustõlke Keskus. ca 1200 märksõna. (http://www.legaltext.ee/; http://www.mig.ee/sonastik/),
- Viiekeelne Euroopa Liidu sõnastik. Eesti-inglise-prantsuse-saksa-soome (1998). Originaal: Europan unionin sanasto. Tõlkinud ja toimetanud Küllike Maurer, Raivo Rammus, Hille Saluäär. Eesti Õigustõlke Keskus. (http://www.legaltext.ee/),
- Andrus Mölder. Väike majandusterminite seletav sõnastik (2002). 236 kirjet. Majanduskonsultatsioonide OÜ. (http://www.mkonsult.ee/Majandusterminid.htm)
- Kaupo Suviste. Windowsi kasutajaliidese põhiterminite kuuekeelne lühisõnastik. 230 kirjet. (http://my.tele2.ee/aidu2/UFF/uff.html),
- Õigusmõistete tesaurus (1999). Koostanud Ene Vainik. Eesti Õigustõlke Keskus. (http://www.legaltext.ee/). Koostatud WordNeti (vt. eespool) printsiipidel.

Sõnastikuosad muudes raamatutes:

- Arne Ansper. Turvaline elektronpost (1996–98). Küberneetika AS. (http://www.cyber.ee/infoturve/ressursid/smail/09.html),
- Vello Hanson. Infosüsteemide turve, 1. Turvarisk (1997). Küberneetika AS. (http://home.cyber.ee/vello/osa1.html). Raamat ja sõnastikuosa pdf-formaadis.
- Teet Jagomägi. Kaardid ja GIS 2000. aasta rahvaloendusel (1997). AS Regio (http://www.geo.ut.ee/gis2000/terminid.html),
- Arne Ansper. Turvaline elektronpost. Küberneetika AS 1996–98 (1999). (http://www.cyber.ee/infoturve/ressursid/smail/09.html).

Siinne oskussõnastike nimestik ei ole kindlasti mitte ammendav. Väikesõnastike või asutusesiseste terminoloogiaandmebaaside olemasolu fakt jäetakse kahjuks sageli ainult enda teada. Kuigi sõnastikku võiks tarvis minna teistelgi kui ainult ühe asutuse inimestel. Enamasti on nad ilmselt siiski väga majasisese iseloomuga, pidevalt poolikud ja üldse mitte avaldamiskõlbulikud.

Et oskussõnavara arendamine nõuab paljude asutuste ja asjatundjate ühist jõupingutust, oli Eesti Keele Instituut üks terminoloogiaühingu algatajaid. Eesti Keele Instituudi ja Eesti Õigustõlkekeskuse 16. märtsi 2000. a nõupidamise tulemusena otsustati luua Eesti Terminoloogia Ühing (ETER).

Tiiu Erelt on koostanud eesti oskussõnastike nimestiku aastatest 1991 – 1999 (http://www.eki.ee/keeleabi/artiklid2/oskuss.html). Seal on kirjas 153 sõnastikku. Nimekirjas olevad sõnastikud on ebaühtlase kvaliteediga ja samuti pole teada, kas neil on olemas arvutiversioon.

2.3.3. Kaks- ja mitmekeelsed sõnastikud ja andmebaasid

Kirjastuste kogemused on, et elektroonilisi sõnastikke vajatakse pigem kaks- kui ükskeelseid ja enamasti tuntakse huvi eriti mahukate sõnastike vastu. Ja mahukas mitte niivõrd sõnaseletuste kuivõrd just tähenduserinevuste, mõnede täpsustavate näitelausete jms. poolest. Samuti ollakse kindlasti huvitatud häälduses heli kujul, mis on aga väga kulukas töö. KeeleWebi'lt küsitaksegi kõige rohkem kakskeelseid tõlkesõnastikke.

Järgnevalt nimekiri arvutis olevatest kaks- ja mitmekeelsetest sõnastikest:

- Eesti-vene sõnaraamat I–V (1997–; tegemisel: 3. köide ilmub 2003). Umbes 60 000 märksõna. Sõnamuutmise infoga on varustatud kõik (eesti) märksõnad (see info lisatakse automaatselt) ja kõik (vene) tõlkevasted. EKI-s olemas andmebaas.
- Vene-eesti sõnaraamat I–IV (1984–1994). 75 000 märksõna. EKI-s olemas nii struktuurimärgenditega lihttekst kui ka andmebaas.
- Vene-eesti-vene online-sõnastik 1997. (http://www.ase.ee/dict/dict.html). EKI vene-eesti sõnastiku veebiversioon. Sõnastiku põhisuund on tõlkimine vene keelest eesti keelde. Tõlkimine vastassuunas (eesti-vene) võib olla ebatäpne.
- EtJaTik eesti-jaapani sõnastik – eesti jaapani sõnaloend, esialgne versioon. Sõnaloend on genereeritud jaapani-eesti leksikograafilisest XML baasist. (http://www.zone.ee/jatik/).

- Eesti-inglise-saksa-soome-vene sõnaraamat – Kaarel Kaldvee. Sõnastik koosneb 700 sõnast paralleelselt kõigis keeltes. (http://www.hot.ee/1955/sonaraamat).
- Inglise-eesti sõnastik – EKI. Eesmärgiks on koostada uus inglise-eesti sõnaraamat, mis oleks vabavara ning orienteeritud ennekõike arvutis kasutamisele. Valmis on inglise märksõnastik, eesti vasted on põhjalikumalt lisatud A-haunt (u. 70000). Andmebaas SGML-märgendusega. (http://kiisu.eki.ee/).
- Inglise-eesti-inglise sõnaraamat (http://www.dict.ee/).
- Baski-eesti sõnastik – Koostanud Sven-Erik Soosaar, sõnastikus u. 800 sõna. (http://lepo.it.da.ut.ee/~svenerik/baski.html).
- Eesti-hollandi sõnastik – Sõnastikus on u. 3500 sõna, valimik kohanimesid, numbrid, kuude ja nädalapäevade nimetused. (http://www.eki.ee/dict/estned.html).
- Eesti-inglise kaupade ja teenuste nimetuste sõnastik – Patendiamet. Umbes 700 üldtehnikasse kalduvat terminit, liigitus nii tähestiku alusel kui ka tootegruppide kaupa.
- Eesti-inglise sõnastik – Aare Vesi. Sisu esialgu sama, mis EKI inglise-eesti sõnastikul.
- ESTERM – Õigustõlke keskuse terminibaas, 27000 inglise-eesti paari (http://www.legaltext.ee/et/andmebaas/ava.asp?m=032).
- Karjala-vepsa-soome-eesti sõnastik – Vepsa selts, Leo Baskin; aegamööda täienev sõnastik sisaldab esialgu u. 1500 sõna neljas keeles. (http://www.veps.de/Sanasto/).
- Ladina-eesti sõnastik – Sõnastik on ilmunud paberkujul 1955. a. A. Härma Ladina keele õpiku lõpus ja sisaldab ca 2000 märksõna. (http://www.ut.ee/klassik/dict/).
- Ladina-eesti väljendeid – Erkki Laaneoks. Lisaks mõned keelelised tähelepanekud ja tarkvara. (http://kodu.neti.ee/~qj223a/).
- Matemaatikasõnastik – 9-keelne matemaatikasõnastik põhiliselt koolimatemaatika terminitest. Keeled on: läti, inglise, prantsuse, eesti, vene, saksa, leedu, poola, rootsi. (http://www.lanet.lv/miv/).
- Mitmekeelsete tõlkesüsteemide kasutajaliidesed – Sõnastikud, päringud sõnastikest, võimalus tõlkida tekste ja veebilehekülgi. (http://www.teataja.ee/sonastik.html).
- Sõjandusterminoloogia – Kaitseliit on välja pannud paarsada eestikeelset oskussõna. Seletusele lisaks on sama sõna vasted

inglise, saksa, vene ja soome keeles. (http://www.kaitseliit.ee/Text/Terminid/).
- Soome-eesti suursõnaraamat I–II (EKI, Kotimaisten kielten tutkimuskeskus, Helsinki; lõpetamisel). 90 000 märksõna. EKI-s olemas struktuurimärgenditega (SGML) lihttekst.
- IBS eesti-inglise (http://www.ibs.ee/dict/) (1994). Inglise-eesti sõnastik sisaldab umbes 17000 sõna koos tõlkega.

Järgnevalt on nimekiri erinevate kirjastuste müügil olevatest sõnastike CD-versioonidest (andmed võetud internetist 2003):
- Eesti-inglise sõnaraamat 2001 + CD. FESTART. Toimetajad Mari Kerge, Maarja Märss, Inga Mölder. 65 000 eesti märksõna. Selle aluseks on 1999. aastal trükis ilmunud sama firma eesti-inglise sõnaraamat.
- FESTART Dictionary: Inglise-eesti-inglise sõnaraamat ver. 3.30 PRO CD (2002). 115 000 inglise märksõna ja 130 000 eesti märksõna, 30 MB kõvakettal. Kokku pandud 2001 ilmunud eesti-inglise sõnaraamatu ning eesti-inglise majandusssõnastiku baasil. Versioonis 3.30 PRO on lisatud uus funktsioon – sõna laiendatud otsing. Kui tavaline otsing leiab ainult konkreetselt sisseviidud sõna või väljendi, siis "sõna laiendatud otsing" annab võimaluse leida selle sõna esinevuse terve sõnaraamatu märksõnade/väljendite hulgas.
- FESTART Dictionary: Vene-eesti-(vene) sõnaraamat versioon 3.27 CD (2002). 47 000 vene märksõna, 6 MB kõvakettal.
- FESTART Dictionary: Inglise-läti-(inglise) sõnaraamat versioon 3.25 CD (2000). 45 000 inglise märksõna, 9 MB kõvakettal.
- Runner Lite. Eesti-vene, vene-eesti sõnastik REAL SOFTWARE. Sõnavara maht on 110 000 sõna. See on ettenähtud tõlkimistööde kergendamiseks ja kiirendamiseks.
- Runner Lite Net. Eesti-vene, vene-eesti sõnastik REAL SOFTWARE. See on Runner Lite programmi võrkversioon ja toetab kõike lokaalse versiooni võimalusi ning on ette nähtud ettevõtetele ja õppeasutustele.
- Adverb 2000. Eesti-vene, vene-eesti tõlkesüsteem REAL SOFTWARE. Sõnavara maht on 175 000 sõnavormi. Sisaldab

sõna tõlget kolmes peakäändes ja tunneb sõna enamikku grammatilisi vorme. See on ettenähtud mustandtõlkimistöö jaoks.
- Eurotranslator.net. Saadaval 21 keelt, sealhulgas ka eesti keel. Koostanud soome firma Sandstone.
- Euroword Software'i sõnastik Euroword 99 sisaldab soome-eesti sõnastikuosa.

Festart'il on plaanis koostada ka saksa-eesti-saksa sõnaraamat, kuid momendil pole see töö firmale majanduslikult otstarbekas.

2.3.4. Andmekogud

Lisaks sõnastikele on mitmeid elektroonilisi andmekogusid, mida saaks kasutada uute sõnastike koostamisel (vt. ka Langemets 2002):
- Sõnaloendid.
  a. Lemmade (ehk algvormis sõnade) loend. Koostatud EKI-s eri sõnastike alusel (k.a. ÕS 1999), sisaldab ca 100 000 sõna.
  b. Vormide (ehk muutevormis sõnade) loend. Koostatud EKI tekstikorpuse alusel, sisaldab ca 200 000 muutevormi (sõnaühendeid pole).
  c. Loendid internetis EKI kodulehelt (http://www.eki.ee/tarkvara/wordlist/) (sõnaloendeid (eriti koos statistiliste andmetega) läheb vaja uute sõnastike tegemisel).
- Kohanimeandmebaas (KNAB). Üle 70 000 kirje. Otsimootor internetis EKI kodulehelt (http://www.eki.ee/knab/knab.htm). Eesti kohanimesid on 32 000 ja väliskohanimesid 39 000.
- Eesti rahva nimed. 138 000 perekonnanime, 52 000 eesnime. Otsimootor internetis KeeleWebi leheküljel (http://ee.www.ee/Nimed/). Andmed pärinevad Eesti rahvastikuarvestuse andmebaasist (andmebaasi peab üleval AS Andmevara) 1995. a sügise seisuga. Avalikus kasutuses on ees- ja perekonnanimed, mida esineb Eesti rahvastiku hulgas rohkem kui 5 korda. Vastused saadakse koos sagedusandmetega.
- Nüüdisvõõrsõnade andmebaas (VSAB). Tegemisel: oktoober 2002 seisuga on baasis 7000 kirjet. Sisaldab väga mitmekesist

infot uuemate võõrsõnade kohta. EKI-s arvutis struktuurimärgenditega lihttekst.
- Uute sõnade andmebaas (USS). Tegemisel: 2001. aasta lõpu seisuga 10 000 märksõna. Koondab erisugust infot eesti uuema sõnavara kohta. Andmebaasi põhjal koostatakse "Eesti kirjakeele seletussõnaraamatu" täiendvihik, mis ilmub koos viimase põhivihikuga aastal 2007. Valmis kirjeid on esialgu vähe. EKI-s arvutis struktuurimärgenditega lihttekst.
- Püsiühendite andmebaas (2002). 17 493 kirjet. Allikteosed: "Fraseoloogiasõnaraamat" (Õim 1993), "Eesti kirjakeele seletussõnaraamat", Filosofti tesaurus, partikkelverbide loend "Das Estnische Partikelverb als Lehnübersetzung aus dem Deutschen" (Hasselblatt 1990), "Eesti keele mõistelise sõnaraamatu" indeks (Saareste 1979), Sünonüümisõnastik" (Õim 1991), tekstist sõnade koosesinemiste (kollokatsioonide) tuvastamise programmi SENVA (Software for Extracting N-ary Verbal Associations) abil 1990ndate aastate korpustest leitud fraasid (Kaalep, Muischnek 2002a, 2002b). Internetis TÜ arvutuslingvistika kodulehel. (http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html).

2.4. Eesti keeletehnoloogia ja lingvistika vajadused elektrooniliste sõnastike osas

Kogu maailmas on tulevikku vaadates arvutisõnastike tegemisel 2 põhisuunda:
- leida uusi lähenemisi elektrooniliste sõnastike tegemiseks nii korpustest kui teistest allikatest, mida saaks kasutada nii keeletehnoloogilistes rakendustes kui ka tavaliste trükitud (või CD-ROM-idel) sõnastike väljaandmiseks;
- leida organisatoorsetele küsimustele uusi lahendusi: erinevate leksikaalsete andmete standardiseerimine, uurimine, arendamine, lahendused sõnastike avalikuks kasutamiseks.

Eesti sõnastike koostajatel on samad eesmärgid (Orav, Muischnek 2002a; Orav, Muischnek 2000b).

2.4.1. Vajalikud leksikonid ja nende ülesehitus

Leksikoni ülesehituse otsustamisel tuleb silmas pidada tervet rida parameetreid, näiteks leksikoni mahtu, info esitusviisi jm. Nende parameetrite väärtusi fikseerides tuleb iga leksikoniprojekti puhul muidugi silmas pidada konkreetseid eesmärke ja vajadusi. Eesti keele kontekstis oleks vaja:

- mahult **suuri** (sõnu sadades tuhandetes) **elektroonilisi sõnastikke**: eesti keele seletussõnaraamat, inglise-eesti, saksa-eesti jm;
- **elektroonilisi sõnastikke**, mis ei ole kättesaadavad ainult kasutajaliidese kaudu, vaid mida oleks võimalik salvestada endale tervikuna (ja kasutada nt. keeletehnoloogilistes rakendustes);
- **korpusepõhiseid läbinisti deskriptiivseid sõnastikke**, kus sõna esinemise sõnastikus määrab ainult tema esinemine korpuses, mitte koostaja subjektiivne arvamus ega eelmiste sõnastike sisu. Siiani on eesti keele tekstikorpused olnud mahult liiga väikesed, et anda adekvaatset infot mingi sõna esinemuse ja tema tegeliku kasutuse kohta. Samuti eeldab see kvalifitseeritud töö-jõu – arvutileksikograafide – olemasolu, kes oleksid suutelised nt. kohandama korpuste töötlemise ja selle alusel leksikonide koostamise programme eesti keelele;
- luua **terminoloogia-andmebaase**; seejuures ära kasutada juba ilmunud (k.a. Nõukogude Liidu ajal ilmunud) oskussõnastikud ja viia need elektroonilisele kujule;
- koostada **eesti keele püsiühendite, kollokatsioonide sõnastikke** (mitte ainult üldkeele, vaid ka allkeelte omi), seejuures ka tõlkesõnastikke;
- **suurendada eesti üldkeele tesauruse e. eesti wordneti mahtu** vähemalt saja tuhande sünohulgani, et see oleks reaalsetes rakendustes kasutatav; samal ajal töötada välja kitsamate valdkondade tesaurusi;
- et **sõnastikud leksikaalsete andmebaasidena** oleksid **arvutis üles ehitatud** nii **paindlikult**, et neist on võimalik teha eri mahu, eri detailsuse, eri raskusastmega uusi sõnastikke. Samuti, et kakskeelsete sõnastike puhul oleks võimalik vahetada lähtekeelt (või esitada neid mõlemas keeles);

- tagada, et **eesti keele kui morfoloogiakeele infoesitust suvalistes sõnastikes toetab automaatne kirjegeneraator.**

2.4.2. Standardiseerimine

Eelnevalt sai räägitud sõnastike standardiseerimise ja korduvkasutatavuse probleemidest mujal maailmas. Eestis on alles hakatud teadvustama seda probleemi. Meil puudub ühtne **struktuurimärgenduse standard** (XML- või SGML-märgendus koos dokumenditüübi definitsiooniga, DTD), mis reguleeriks ja ühtlustaks *kõigi meie sõnastike* andmete esitust. Mõnede sõnastike puhul on seda kasutatud, näiteks soome-eesti sõnaraamatus on kasutatud SGML-märgendust, aga terviklik süsteem puudub. Osa sõnastikke on tehtud lihttekstist ümber andmebaasiks (näiteks Eesti Kirjakeele Seletussõnaraamat TÜ-s). See kogemus näitas, et selline töö on suure aja- ja energiakuluga. Sellise lisatöö aitaks ära hoida just ühtne märgendussüsteem.

Samuti tuleb selle punkti all osutada, et Eestis on väga vähe eesti keele sõnastike tegemist toetavat **tarkvara** (nii on paljude keelte puhul kasutatavad näiteks FrameMaker – SGML-märgenduse süsteem; WordManager). Niisugust tarkvara ei pea spetsiaalselt eesti keele jaoks välja töötama, piisab teiste keelte jaoks loodud tarkvara kohandamisest.

2.4.3. Mitmekeelsus

Puudus on:

- põhjalikust tänapäevasest elektroonilisest inglise-eesti ja eesti-inglise sõnastikust;
- sihtrühma poolest täpsemini määratletud sõnastikest, kui seda on praegu kättesaadavad sõnastikud (nt. kaks täiesti erinevat inglise-eesti sõnastikku, üks inglastele ja teine eestlastele; eraldi sõnastikud eri astmetel õppijatele, eraldi kirjutamiseks ja lugemiseks mõeldud sõnastikud jne.);

- mahukatest ja kvaliteetsetest mitmekeelsetest arvutisõnastikest, mis aitaks teha tõlketöid. Siiani suudetakse täita suhteliselt kitsaid ülesandeid, näit. tõlgitakse õigustermineid. Laiale tarbijaskonnale orienteeritud rakenduste, näiteks internetitekste tõlkivate masintõlkesüsteemide puhul vajatakse tunduvalt laiemat sõnahulka.

2.4.4. Organisatsioonilised aspektid

Siin võib osutada eelkõige kahele aspektile. Esiteks on vaja enne leksikoni väljatöötamisele asumist välja selgitada, missuguses ulatuses on võimalik kasutada olemasolevaid elektroonilisi materjaliallikaid (sõnastikke, korpusi), missuguses ulatuses on võimalik tööd automatiseerida, aga seega ka missugune on prognoositav käsitsitöö maht.

Siia kuulub samuti intellektuaalse omandi õiguse (*copyright*) küsimus. Paljud kirjastused pole nõus sõnastikke vabalt kättesaadavaks tegema just kopeerimiskaitse puudumise pärast. Üheks heaks näiteks on leping, mis sõlmiti Keskkonnaministeeriumi ja erafirma Filosoft vahel 2002. a. Selle lepinguga sai eesti keele spelleri-programm vabalt kasutatavaks, st. riik ostis selle programmi erafirmalt Filosoft. Samamoodi saaks teha ka oluliste leksikaalsete ressurssidega.

2.4.5. Kokkuvõte

Edasine areng kui selline on täies ulatuses ennustamatu, kuid välja võib tuua järgmised aspektid:

- sõnastike masstootmist Eestis CD-de kujul ilmselt ei tule. Need jäävad ikka mingisse lisatoote staatusesse;
- tundub nii, et parimad üldkeele- ja oskuskeele sõnastikud lähenevad omavahel, st. parimate üldsõnastike lõplik esituskuju on üsna sarnane parimate oskussõnastike omaga (koostamismeetod on risti vastupidine ja andmebaas väga erinev, aga tulemus näeb sarnane välja). Ja see lähenemine jätkub;

- sõnastike koostajad hakkavad rohkem pöörama tähelepanu kvaliteedi keskmisele tõusule, st. ühest küljest võib oodata täielike avantüristide jätkuvat kadumist turult, ja teisest küljest praeguste hoolega tehtud asjade edasist parandamist;
- luuakse mahuliselt suuri korpusi, mille andmete analüüsi põhjal saab koostada tegeliku keelekasutusega sõnastikke. Sõnastikke saab panna hulgaliselt tekstinäiteid korpusest, mis annab adekvaatsema pildi sõna tegelikust kohast ja rollist keele struktuuris;
- arvutisõnastikes hakatakse kasutama kõnetöötluse vahendeid – näiteks eesti keele häälduse õpetamisel vene või muud võõrkeelt kõnelevale inimesele;
- teadmusbaasid saavad võimalikuks koos semantilise analüüsi vahendite täiustamisega, st. siis, kui baas oskab end ise täiendada saadavalolevate tekstide põhjal ega sõltu oma arengus ainult inimpsüühikast. Eestis ei oska küll arvata, millal midagi ligilähedast tulla võiks;
- masintõlkesüsteemidega hakatakse tegelema innukamalt. Huvi on ilmutanud Festart ja mõned üksiküritajad;
- põhiline praktiline mure Eestis on **inimressurssid.** Arvutileksikoni loomine – nii nagu iga teinegi arvutuslingvistiline või keeletehnoloogiline rakendusülesanne – eeldab mitme eriala inimeste koostööd. On vaja inimesi, kes oleksid piisavalt kompetentsed keeleteoorias, leksikoloogias ja leksikograafias, keelekirjelduse formalismides, korpuste kasutamises, arvutilingvistikas ja arvutiteaduses. Alles siis, kui on olemas inimesed, kes suudavad ja oskavad leksikaalse infoga adekvaatselt tegeleda igas nimetatud aspektis, alles seejärel saame töötada tõeliselt tänapäevasel tasemel (Langemets 2002).

3. KEELETEHNOLOOGLINE TARKVARA

Alljärgnevalt kirjeldame arvutitarkvara, mida kasutatakse keele ja kõne töötlemisel, olgu see siis tavalises kontoris ja kodus (kus keelt ja kõnet käsitletakse eeskätt kui informatsiooni edastamise vahendit), või teadustöös (kus keel ja kõne on ise uurimisobjektid). Algul kirjeldame tarkvara kasutusvaldkondi ja seda, mis maailmas neis valdkondades on tehtud; seejärel seda, mis Eestis on valdkonniti olemas, ja lõpuks seda, mida Eestis oleks vaja teha.

3.1. Tarkvara kasutusvaldkonnad

3.1.1. Lõppkasutajale mõeldud programmid

3.1.1.1. Kirjutaja abivahendid

Kirjutaja abivahendite alla kuuluvad mitmed laialt kasutatavad rakendused. Üks elementaarsemaid on sõnade automaatne poolitus, nt. *plekk*/*trumm*, mitte *pekkt*/*rumm*. Poolitus tugineb keele häälikuu- ja silbistruktuuri ning sõnavara tundmisele. Õigekirja kontroll omakorda tugineb suurtele sõnaraamatutele, sõnamuutmisreeglitele, liitsõnade ja tuletiste moodustamise reeglitele, grammatikareeglitele ja suurtele tekstikorpustele. Kontrolli käigus leitud vigade parandamine tugineb lisaks veel teadmistele tõenäolistest veatüüpidest. Siia valdkonda kuulub ka võimalus pöörduda tesauruse vm. sõnastiku poole otse tekstis oleva sõnavormi peal klõpsates, ilma et peaks leidma enne algvormi, nt. *poest* peale klõpsates saame vajadusel vasteks sõna pood sünonüümid õiges vormis, nt. *kauplusest*.

Kirjutaja abivahendid on kõige vanem ja levinum keeletehnoloogia valdkond, aga valdkonna kõiki võimalusi pole veel kaugeltki realiseeritud.

3.1.1.2. Dokumenditöötlus

See on lai valdkond, kuhu kuulub mitmeid keeletehnoloogia komponente. Neist olulisemad on info-otsimine, dokumendi keele tuvastamine, dokumentide liigitamine, kokkuvõtete tegemine, hüperteksti ja viitade automaatne genereerimine, terminikogude loomine ja terminoloogide abivahendid.

Info- ja dokumendihalduses ning infootsingus kasutatakse enamasti üksikute sõnade töötlemiseks sobivaid keeletehnoloogilisi võtteid, mis on osalt samasugused kui kirjutaja abivahendeis. Kuna suur osa info-otsimise meetoditest ja tarkvarast on pärit inglise keelt kõnelevatest maadest, siis ei ole neis tüüpiliselt arvestatud probleeme, mis tekivad sõnade käänamisest ja pööramisest, liitsõnade moodustamisest ja sõnatuletusest. Need probleemid on omased just rikka morfoloogiaga keeltele nagu eesti, soome, türgi jms. Nende ignoreerimine muudab otsimise ebatäpsemaks, nt. otsisõna *pood* ei võimalda leida vorme *poes* ja *poest*. Keeletehnoloogia pakub siin mitmeid lahendusi, mida saab kasutada nii täpsete kui hägusate (*fuzzy*) otsimismeetodite puhul. Morfoloogilise analüüsi abil on võimalik leida sõnade algvormid ja liitsõnade osasõnad ning koostada neist indeks, mida info-otsimisprogrammid kasutavad. Nii saab päringuga *katus* otsida ka *plekkkatust*. Teine võimalus on moodustada kasutaja antud päringusõnast kõik sõnavormid ja siis neid kõiki tekstidest otsida, s.t. tuleb kasutada morfoloogilist sünteesi. Probleemiks on mõlemal juhul see, et nii sõnad kui nende vormid on mitmeti tõlgendatavad, millest üle saamiseks tuleb arvestada ka sõnade konteksti.

Dokumente on võimalik neis sisalduvate sõnade alusel ka liigitada. Seejuures kasutatakse valdavalt statistilisi meetodeid, mis on algselt mõeldud inglise keele jaoks, kuid mida saab kasutada ka morfoloogiliselt keerulisemate keelte korral, kui sõnad algul viia

algvormide kujule. Kui võrrelda kahte dokumenti neis sisalduvate sõnade esinemissageduste põhjal, siis algvormide põhjal tehtud statistika annab enamasti parema tulemuse kui sõnavormide peal tehtu. Nt. laused *Parlamendis vaieldi valitsuse eelnõu üle* ja *Valitsus pani oma eelnõuga parlamendi vaidlema* kirjeldavad tõenäoliselt sama asja. Seda on ilmselt raske automaatselt leida, kui sõnad pole enne taandatud algvormideks.

Samal moel võib läheneda ka dokumentidest kokkuvõtete tegemisele, kus proovitakse automaatselt eristada need dokumendi osad, kus arvatakse olevat dokumenti kõige täpsemalt iseloomustav tekst. Raamatust indeksisse minevate terminite automaatne valik on samuti juba vana ülesanne, mida saab lahendada automaatselt. Uusim terminite äratundmise kasutusala on automaatne linkide tekitamine hüperteksti. Viimatinimetatud kasutusala puhul on õieti tegemist tähendustega; kuid tehniline lähenemine annab küllalt täpseid tulemusi, kui vastab tõele eeldus, et sõnade esinemine peegeldab tähenduste esinemist.

Keeletehnoloogiat saab info-otsimises kasutada ka sel moel, et kasutatakse olemasolevaid sünonüümisõnastikke või tesauruseid alam- ja ülemmõistetega, et otsitakse tekstist sageliesinevaid ja püsivaid väljendeid, et eraldatakse lause analüüsi käigus väljendite keskseid osad vähemtähtsatest või kasutatakse konteksti, et mitmetitõlgendatavust vähendada.

Keeletehnoloogiliste rakenduste hulka kuulub ka ükskeelsete ja mitmekeelsete sõnastike kasutamine päringute tegemisel. Ükskeelsed sõnastikud, nt. sünonüümisõnastik ja WordNet-tüüpi andmebaas pakuvad variante, mida lisaks esialgsele otsi-terminile kasutada. Nii saab sõnastikku kasutada algse päringu automaatseks või pool-automaatseks laiendamiseks, mis võib oluliselt otsimistulemust parandada. Mitmekeelse sõnaraamatu abil saab ühendada info-otsimist mitmetest erikeelsetest dokumendikogumikest. See on just viimasel ajal muutunud oluliseks uurimisalaks ja selle tähtsust tõstab EL-i laienemine.

3.1.1.3. Tõlkijate abivahendid ja lokaliseerimine

Lisaks arvutile toetuvatele tõlke-abi programmidele kuuluvad siia alla mitmesugused töövahendid terminoloogia haldamiseks ja elektroonilised sõnaraamatud. Need võivad olla nii tavalised infotehnoloogilised vahendid (nt. andmebaasi haldamise tarkvara) kui ka programmid, mis sisaldavad keeletehnoloogiat, s.t. spetsiaalseid just inimkeele töötlemiseks mõeldud funktsioone (nt. täisteksti andmebaasi kasutamisele ülesehitatud rakendus, milles sõnaindeksi tegemisel teisendatakse tekstis esinevad sõnad algvormi kujule). Tõlkimine on nt. Euroopa Liidu valitsusala suurim kuluartikkel, mis annab tööd tuhandetele tõlkidele ja tõlkijatele. Arvutile toetuvad tõlkeprogrammid jagunevad laias laastus sõnastiku- või tõlkemälu-põhisteks tõlke-abi programmideks ja nn. tõelisteks masintõlkeprogrammideks. Sõnastiku-põhised oskavad pakkuda lähtetekstis olevale sõnale konteksti sobivat vastet sihtkeeles. Tõlkemälu-põhised säilitavad mälus varem tõlgitud lähte- ja sihttekstid. Uut teksti tõlkides otsitakse mälust võimalikult sarnane varemtõlgitud tekstiosa ja pakutakse selle varemtehtud tõlget ka uude tõlkesse. Nn. tõelised masintõlkeprogrammid tõlgivad varem mitte kohatud lauseid juba täies ulatuses.

3.1.1.4. Masintõlge

Seni on masintõlke suurim puudus võrreldes inimese tehtud tõlkega see, et masin ei saa tekstist aru ega tea, mille jaoks tõlget tehakse. Tänapäeval võib eristada 3 levinumat lähenemisviisi masintõlkele; neid püütakse omavahel ka kombineerida.

1. Transfer-meetod. Arvuti teisendab teksti osalausete kaupa, grammatikat ja kakskeelset sõnastikku kasutades. Algul tehakse lähteteksti morfoloogiline ja süntaktiline analüüs, siis teisendatakse saadud struktuure ja viiakse nad teisele keelele omasemale kujule; seejärel otsitakse leksikonist sõnade ja väljendite vasted ning lõpuks tehakse süntaktiline ja morfoloogiline süntees. Sellest tuleneb, et tõlge on parimal juhul üsna sõna-sõnaline. Selleks, et masin oskaks paljudest võimalikest tõlkevariantidest valida

konteksti sobivat, tuleb teda reguleerida ehk tema grammatika ja sõnastik sobitada tõlgitava teksti tüübi ja valdkonnaga. Teksti mõistmiseks tuleb mõnikord lähtekeele kompaktne teade jagada üksikuteks väideteks ja moodustada neist väljundkeelele omane liitlause. Seda tänapäeva masintõlkesüsteemid ei suuda, vähemalt mitte eriti loovalt.

2. Tõlkemälu meetod (ingl. k. *translation memory*). Suur hulk tõlkimist vajavatest tekstidest on juhendid, eeskirjad, käsiraamatud jms., kus kasutatakse standardseid väljendeid ja millest sageli antakse välja uusi, kaasajastatud versioone. See tähendab, et ka nende tõlked sisaldavad palju standardseid väljendeid ja erinevad varasematest versioonidest vaid üksikutes kohtades. Seega saab suure osa tõlketekstist panna kokku varemtõlgitud tekstide osadest. Kui meil on tõlkemälu e. lähtekeele blokid koos neile vastavate tõlgetega, siis saab uue teksti puhul otsida sealt juba varem tõlgitud blokke ja need uude teksti otse asendada. Probleemiks on tõlkemälus säilitatavate blokkide suuruse valik. Mida suurem on mälus olev blokk, seda suurem on tõenäosus, et ta on adekvaatne tõlge, kuid seda väiksem on tõenäosus, et see blokk üldse mõnes uues tekstis esineb. Bloki suurust vähendades suurendame tõenäosust, et ta uutes tekstides esineb, kuid samas vähendame ka tõenäosust, et tema asendamine uutesse tekstidesse annab õige tõlke.
3. Müranivooga kanali (ingl.k. *noisy channel*) meetod. Oletame, et algselt taheti tekst kirja panna selles keeles, millesse me tahame teda tõlkida, aga mingi müraallikas moonutas kirjapandut, nii et nüüd on ta teises keeles (nt. „Hamlet“ pidi olema algselt eesti keeles, aga Shakespeare pani ta kirja inglise keeles). Tuleb leida viis, kuidas taastada teksti „originaalne“ e. algselt mõeldud kuju. Sellise lähenemise korral saab kasutada mitmeid informatsiooniteooriast, side, kommunikatsiooni, raadio, kõnetuvastuse jm valdkondadest pärit algoritme. Ehkki lähenemine võib tunduda imelik, on see tegelikult samasugune kui kõnetuvastuse probleemipüstitus: loomuliku keele lauset on moonutatud nii, et temast on saanud akustiline signaal, ja ülesandeks on leida esialgselt mõeldud lause.

Olgu **P(e|i)** tõenäosus, et eestikeelne lause **e** oli ingliskeelse lause **i** originaal. Lähtudes olemasolevast ingliskeelsest lausest **i**, taandub masintõlke probleem sellele, et tuleb leida selline eestikeelne lause, mille puhul tõenäosus **P(e|i)** on suurim. See tähendab, et me otsime $\hat{e} = \mathbf{argmax}_e \mathbf{P(e|i)}$. Bayesi teoreemi kohaselt $\hat{e} = \mathbf{argmax}_e \mathbf{P(e|i)} = \mathbf{argmax}_e \mathbf{P(i|e)\ P(e)}$. **P(i|e)** on tõenäosus, et **e** tõlkimisel saadakse **i**. **P(e)** on tõenäosus, et lause **e** üldse eesti keeles võib esineda.

Tõenäosused leitakse paralleelkorpuste, s.t. korpuste, milles on tekstide originaalid ja nende tõlked, alusel.

Võrdlustestid on näidanud, et müranivooga kanali meetod võimaldab teha sama kvaliteediga masintõlget kui transfer-meetod.

3.1.1.5. Keeleõpe arvuti abil ja sõnaraamatud

Seni on keeleõppeprogrammides ja elektroonilistes sõnaraamatutes kasutatud enamasti traditsioonilist infotehnoloogiat, kuid võiks lisada ka keeletehnoloogilisi rakendusi.

Keelekursustel pööratakse sageli põhitähelepanu õpitava keele sõnavara ja grammatika õpetamisele, korrektse häälduse õpetamine on sageli ebapiisav. Kasutades kõnesignaalide analüüsivahendeid on võimalik luua võõrkeelse häälduse hindamise ja treenimise süsteeme, mis koos vastava treeningmetoodikaga aitavad omandada võõrkeelt aktsendivabalt.

Elektrooniliste sõnaraamatute uued põlvkonnad juba sisaldavad keeletehnoloogia elemente (algvormide leidmist, fraaside automaatset leidmist tekstist jm.).

3.1.1.6. Kõnesüntees

Kõnesüntees tähendab (elektroonilise) teksti teisendamist kuuldavaks kõneks.

Suur kasutajate ring on mitmesuguste puuetega inimesed, põhiliselt pimedad, kelle suhtlemis- ja tööprobleemide lahendamisele aitaksid kaasa mitmed keeletehnoloogia vahendid. Pimedad saavad kasutada kõnesünteesi programmi, mis loeb ette arvutis leiduvat ja internetist saadavat tekstikujulist informatsiooni. Kõnesüntesaator ja skänner koos optilise tekstituvastuse programmiga moodustavad lugemismasina, mis loeb ette paberkandjale trükitud teksti.

Kõnesüntesaator on vajalik ka kõnepuuetega inimestele suhtlemiseks tavainimestega (enamik inimesi ei oska viipekeelt).

3.1.1.7. Kõnetuvastus

Automaatse kõnetuvastuse ülesandeks on mikrofoni kaudu arvutisse sisestatud kõnesignaali teisendamine tekstiks.

Heatasemelise kõnetuvastuse olemasolu, mida ennustatakse lähitulevikuks (5 aasta perspektiivis), võib põhjalikult muuta inimese ja arvuti vahelist suhtlemist, aga ka inimeste omavahelisi suhtlemiskanaleid.

Juba praegu rakendatakse kõnetuvastust mitmesuguste seadmete suuliseks juhtimiseks. Mitmete keelte jaoks on olemas praktiliselt kasutatavad automaatsed diktofonid, s.t. arvuti teisendab kõneldud jutu tekstiks. Kõnetuvastust kasutavad dialoogsüsteemid on leidnud kasutamist kindlalt piiritletud valdkondades, näiteks infootsing telefonikataloogist, lennukite ja reisirongide sõiduplaanidest ning piletite reserveerimine. Reaalselt töötavad sellised infosüsteemid Inglismaal (British Airways lendude info ja piletite reserveerimine), Prantsusmaal (telefoni kataloogi kollaste lehekülgede info), Saksamaal (reisirongide info ja piletite reserveerimine), Itaalias (reisirongide info ja piletite reserveerimine), Hollandis (teatripiletite reserveerimine), Rootsis (Stockholmi turismiinfo).

Kurtide inimeste kõnelema õpetamisel on palju abi kõneanalüüsi vahenditest, mis esitavad erinevaid kõneparameetreid visuaalselt arvutiekraanil. Kurtide inimeste kõnevõime on piiratud just selle tõttu, et neil puudub kõneproduktsiooni juhtiv akustilise tagasiside kanal. Seetõttu on nende kõne ebaloomulik ja sageli tavainimesele arusaamatu. Kui kurtidele esitada nende kõnest leitud parameetrid koos lubatud muutumispiiridega visuaalselt, siis on neil võimalus õppida paremini koordineerima oma kõneorganite tööd ja siis muutub ka nende kõne tavainimesele arusaadavamaks.

Viimastel aastatel on see valdkond arenenud kiiresti ja on muutunud järjest populaarsemaks. Üks põhjusi on muidugi inimeste jaoks loomuliku suhtluskanali – suulise kõne – võimaldamine. Teine on see, et kõnetuvastus keeletehnoloogilise probleemina haarab tegelikult kõiki olulisi keeletasandeid, mitte ainult häälikuanalüüsi: ka morfoloogilist analüüsi (et kindlaks teha, kas väljapakutud sõnavorm on antud keele sõnavorm); süntaktilist analüüsi (kas tuvastatud sõnavormide järjend on antud keele fraas või lause). Seetõttu on keeletehnoloogid leidnud kõnetuvastussüsteemide loomises valdkonna, mis ühendab keele erinevate aspektide modelleerimise kallal töötavate uurijate jõupingutusi, seejuures nii, et lõpptulemused on mõõdetavad ja samal ajal suure sotsiaalmajandusliku ja –poliitilise väärtusega.

3.1.1.8. Loomulikku keelt võimaldavad kasutajaliidesed

See on omaette lai valdkond, peamiselt seotud sellega, et inimene saaks (andmebaasi) päringuid esitada loomulikus keeles. Loomuliku keele kasutamist infopäringul on maailmas küllalt põhjalikult uuritud. Mitmete piiratud valdkondade (sõiduplaanid, piletite tellimine, jne.) puhul on leitud tüüpilised dialoogistruktuurid, mida rakendatakse edukalt ka automaatsetes süsteemides. See valdkond kombineerub kõige otsesemalt kõnetuvastuse ja –sünteesiga (vt. käesoleva ülevaade p. 3.1.1.6. ja p. 3.1.1.7), nii et tulemuseks on telefonipõhised automaatsed infosüsteemid.

3.1.2. Teadus- ja arendustöödeks mõeldud programmid

3.1.2.1. Sõnavormide analüüs

Keeletehnoloogilised võtted, mida kirjaliku teksti töötlemisel kasutatakse, ühendavad suurt hulka keelega seotud teadmisi ja matemaatikat, nt. automaatide teooriat. Kuna inimkeel ei ole mingi väike ega lihtne käsitlusobjekt, on vaja kasutada küllalt võimsaid vahendeid. Nt. eesti keele lihtsõnu sisaldav Ülle Viksi „Väike vormisõnastik" võimaldab 35 000 sõna põhjal moodustada miljon sõnavormi, kui lisada aga võimalikud liitsõnad ja tuletised, siis ulatub võimalike sõnavormide arv miljarditesse.

Sõnavormide analüüs on osutunud siiski piiritletud ja üsna hästi lahendatavaks ülesandeks. Praeguseks on sõnavormide analüüs realiseeritud sõnastike ja reeglite kombineerimise teel. Kuigi sõnavorme on palju, on nende moodustamine küllalt reeglipärane. Sõnavormide analüüsiprogrammi ehk morfoloogilise analüsaatori tegemine on mõne inimaasta suurune töö.

3.1.2.2. Grammatiline analüüs

Kirjaliku teksti lauseliikmete määramine või muu mitut sõna haarav käsitlemine eeldab üldiselt *morfoloogilist analüüsi*. Sõnavormid on sageli mitmetähenduslikud, mis jääb grammatikate koostajatel tihti kahe silma vahele. Nt. ingliskeelne sõna *left* võib olla nii tegusõna kui omadussõna. Eesti keeles *mees* võib olla nii *mees* ainsuse nimetav kääne kui *mesi* ainsuse seesütlev. Sõnavormide homonüümiat esineb kõikides keeltes, aga eri ulatuses. Eesti keeles on umbes 40–50% tekstis esinevatest sõnavormidest mitmeti mõistetavad. Mitmeti mõistetavate sõnavormide hulgast õige valimine ehk morfoloogiline ühestamine on üsna hästi lahendatav probleem, kui arvestatakse sõnade naabrust ehk konteksti. Nt. piisab inglise keele puhul sõnapaarist *he left*, et otsustada, et tegu on tegusõnaga, või eesti keeles *suur mees*, et teada, et tegu ei ole *meega*. Ühestamist võib teha nii statistiliste meetoditega, nt. MVM (Markovi varjatud

mudel, ingl. k. HMM, st. Hidden Markov model) kui ka reeglite abil. Ehkki pole teada ühtegi ühestajat, mis leiaks ainsa sobiva analüüsi kõigile sõnavormidele, on mitmete keelte jaoks olemas ühestajaid, mis suudavad seda anda 95% tekstis esinevatele sõnavormidele.

Süntaktilist analüüsi võib teha eri põhjalikkusega. Kõige pealiskaudsema analüüsi puhul püütakse leida vaid osa sõnade vahelistest suhetest, nt. millised on nimisõnafraasid ja milline on iga fraasi peasõna. Palju keerulisem on luua analüsaator, mille poolt tehtud analüüs on nii täielik, et selle põhjal võib mõista lause sisu või tõlkida lauset teise keelde. Esimest liiki analüüsi võib nimetada *pindanalüüsiks*, teist aga *süvaanalüüsiks*. Pindanalüüs on keerulisem kui morfoloogiline ühestamine, kuid on siiski suhteliselt hästi tehtav.

Pindanalüüsi puhul jääb lauseehituse mõistmine puudulikuks. Nt. ingliskeelsest lausest *I saw the man on the hill with a telescope* selguks, et ma nägin meest ja et mees või teleskoop oli künkal ja et teleskoop on seotud künka, mehe või minuga. Eri tõlgendused annaksid erinevad tõlked eesti keelde. Kõige pindmisem süntaktiline analüüs leiaks lausest neli nimisõnafraasi: *I, the man, on the hill* ja *with a telescope*. Veidi sügavama analüüsi puhul määratakse esimene fraas subjektiks, teine objektiks ja mõlemast fraasist leitakse fraasi peasõna.

Pindanalüüsi saab kasutada muu hulgas õigekeele kontrollimiseks (grammatika kontrollija), terminoloogi töövahendina, keeleõppeprogrammides, info-otsimisprogrammides, aga ka parandada muude, sõna tasemel töötavate programmide kvaliteeti.

Üldtuntud ja põhjalikult läbiuuritud grammatikateooriate (transformatsiooniline grammatika, LFG, GPSG, HPSG, GB) eesmärgiks on lause nii peen analüüs, et selle põhjal saaks mõista lause tähendust ja sellega seoses ka sõnadevahelisi suhteid ja mõjusid. Nii täielikku grammatilist analüüsi on siiski ülimalt raske realiseerida täiesti ühetähenduslikult, kui ei võeta appi semantikat ja situatsioonikonteksti. Lahendamata jäävad terved konstruktsioonitüübid, nt. *saabusid tema välismaal elavad vanemad ja õed*, sest siit me ei saa teada, kus õed elavad. Seega tekst ei sisalda alati õige tähenduse või tõlke valimiseks vajalikku informatsiooni. Sellise taseme lauseana-

lüüsi on uuritud ja arendatud muis mais juba aastakümneid erinevate lingvistiliste teooriate alusel, kuid veel ei paista silmapiiril lõplikku lahendust ega isegi ühtegi üldtunnustatud teed lahenduse suunas. Mõnede meetodite puhul tekib palju alternatiive, mõnede puhul ei kata grammatika tekstis ette tulevaid lauseid küllaldaselt ja mõned analüsaatorid oletavad, tehes seejuures vigu.

Kõrgekvaliteedilise süntaktilise analüsaatori puudumine on tähelepanuvääriv asjaolu, sest selle abil oleks võimalik luua küllalt hea masintõlkesüsteem. Paljud muudki keeletehnoloogilised probleemid laheneksid kvaliteetse süntaktilise analüüsi olemasolul.

Tuleb järeldada, et põhjus on konkreetseid keeli puudutavate süntaksialaste baasuuringute ebatäiuslikkuses – vaatamata teoreetiliste mudelite rohkusele.

Süntaksi analüüsi kasutamine täistekstidel põhinevates info-otsisüsteemides võimaldaks oluliselt tõsta otsingu täpsust, kõnesünteesis aga kõne loomulikkust. Süntaksianalüsaator oleks efektiivne ka tavalistes tekstitöötlusprogrammides, võimaldades automaatselt kontrollida teksti grammatilist korrektsust. Samuti on süntaktiline analüüs eeletapiks teksti semantilisele analüüsile.

3.1.2.3. Semantiline analüüs

Juba lausestruktuuride ühene tuvastamine eeldab sõnade ja neist moodustatud konstruktsioonide tähenduste arvestamist. Näiteks *traktori nahast kate on* süntaktiliselt vähemalt kaheti analüüsitav, kuid semantika praktiliselt välistab analüüsi *((traktori nahast) kate),* jääb *(traktori (nahast kate))*.

Lause- ja fraasisemantika on siiski ka keeleteaduses eneses alles lahendusi otsiv ala, eriti mis puudutab tähenduste formaalset esitamist (mida arvutianalüüs eeldab). Töötavaid lahendusi on olemas vaid kitsamate ainevaldkondade jaoks, mida ka kasutatakse nt. semantiliselt orienteeritud info-otsisüsteemides.

Kuid ka üksiksõnade tähenduste käsitlemine ja arvestamine keeletehnoloogilistes rakendustes võib neile rakendustele oluliselt efektiivsust lisada, ja siin on tulemused märksa paremad. Tuntuim on

nn. leksikaalsete või semantiliste andmebaaside kasutamine nt. dokumendiotsingus või ka dokumentide koostamisel.

Niisuguses andmebaasis ei ole küll sõnade tähendused üheselt defineeritud, kuid on fikseeritud teatud (vastava rakenduse jaoks olulised) tähenduslikud seosed sõnade vahel*:*

sünonüümia e. samatähenduslikkus, hüpo- ja hüperonüümia (*auto* on *liiklusvahendi* üks hüponüüme; *liiklusvahend* on *auto*, *bussi* jne. hüperonüüm), osa-terviku seosed (*mootor* on *auto* üks osi) jne. (vt. ka osa 2. Arvutileksikonid).

3.1.2.4. Pragmaatika

Keele ja kõne analüüs ning süntees ei saa piirduda sõna, fraasi või üksiklause tähendusega. Tegelik keel esineb seotud tekstide, diskursustena. Et teksti üksusi mõista, tuleb arvestada nii tekstuaalset konteksti kui ka tekstivälist maailma, nt. seda, et laused on tekstis teatud viisil (nt. asesõnade kaudu) omavahel seotud, või seda, et viisakas vestluses võib küsimus tähendada hoopis palvet („Kas ma tohiksin soola paluda?“).

Teadmist, milliste vahenditega seotakse tekst või dialoog tervikuks, on vaja nt. automaatsete sisukokkuvõtete tegemisel, loomulikku keelt kasutatavates kasutaja-liidestes, aga ka masintõlkes, kui kasutusel on transfer-meetod.

Sarnaselt semantikaga on pragmaatika ala, kus formaalseid mudeleid ja töötavaid lahendusi on olemas vaid kitsamates rakendusvaldkondades, kus kasutatakse dialoogi ja teksti genereerimist.

3.1.2.5. Kõnesüntees

Kui kõne oleks kirjeldatav nii nagu trükitud tekst – elementaarsümbolite jadana, siis oleks kõne tuvastus ja süntees lahendatud vähemalt paar aastakümmet tagasi. Kahjuks (uurijate õnneks!) on kõne olemus teistsugune: koartikulatsioon liidab häälikud keeruliseks akustiliseks kontiinumiks, kus häälikult-häälikule üleminek sisaldab

sageli rohkem informatsiooni kui häälikud ise. See on ka üheks põhjuseks, miks vaatamata suurtele pingutustele ligi 40-aasta jooksul on suvalise teksti teisendamine arusaadavaks ja loomuliku kõlaga kõneks lõplikult lahendamata probleem ja aktiivse uurimise teema.

Tekst-kõne sünteesi eesmärgiks on teisendada ortograafiline tekst loomuliku kõlaga kõneks. Selleks on vajalikud järgnevad etapid:

- lingvistiline tekstitöötlus: selle tulemusena teisendatakse ortograafiline tekst häälduste kstiks. Sõltuvalt keelest on see erineva raskusega ülesanne (vrd. inglise ja soome keele kirjapilti ja hääldust);
- häälikute kestuse ja lausetüübile (vrd. jutustav ja küsilause) vastava meloodiakontuuri genereerimine;
- kõnesignaali genereerimine: signaali genereerimiseks kasutatava meetodi järgi jagunevad süntesaatorid artikulatoorseteks, formant- ja kompilatiivseteks süntesaatoriteks.

Artikulatoorne süntees baseerub kõneproduktisooni füsioloogilisel mudelil ja kõnetraktis hääle tekkimise füüsikalisel kirjeldusel: modelleeritakse inimese suu, kõri jm. tegevust kõnelemise ajal. Need süntesaatorid on praktiliseks rakenduseks sobimatud, kuna sünteesitava häälelaine arvutamine ei ole teostatav reaalajas. Samas on artikulatoorsetel mudelitel oluline roll kõneproduktsiooni mehhanismide teoreetilisel uurimisel.

Formantsüntees: kõnesünteesi formantmudelid baseeruvad kõnesignaali akustilis-foneetilisel kirjeldusel. Baasmudel koosneb allikast ja filtrist, kusjuures allikas modelleerib häälekurdude võnkumist ja filter kõnetrakti resonantssagedusi – formante. Nii allika kui filtri parameetreid juhitakse erinevate foneetiliste reeglite alusel. Kasutades formantmudelit, on realiseeritud kõrge sünteeskõne kvaliteediga süntesaatoreid erinevate keelte jaoks. Eestis on erinevaid formantsüntesaatoreid välja töötatud Küberneetika Instituudis ja Eesti Keele Instituudis.

Kompilatiivne süntees baseerub naturaalkõnest väljalõigatud signaalilõikude (difoonide, trifoonide, silpide, jm.) sobival ühendamisel. Kompilatiivsünteesiks koostatakse kompileeritavate segmentide andmebaas, mis kajastab enamikku sünteesitava keele fonoloogilisi ise-

ärasusi. Sünteesiprotsessis valitakse andmebaasist sünteesitavale tekstile vastavad segmendid ja ühendatakse spetsiaalse signaalitöötlusalgoritmi abil ühtseks lauseks. Tulemuseks on kõrgekvaliteediline sünteeskõne.

3.1.2.6. Kõnetuvastus

Automaatse kõnetuvastuse ülesandeks on mikrofoni kaudu arvutisse sisestatud kõnesignaali teisendamine tekstiks. Kõnetuvastus on keeruline ülesanne eelkõige kõnelejast tingitud variatiivsuse ja väliste mõjutegurite tõttu:

- häälikud sõltuvad väga tugevasti sellest, millises positsioonis nad sõnas on. Näiteks /n/ sõnas *kong* hääldatakse oluliselt erinevalt võrreldes sõnaga *konn;*
- kõnesignaal sõltub suurel määral keskkonnast ja sidekanalitest, mille kaudu kõnet edastatakse (erinevad mürad);
- erinevate kõnelejate häälduslaneer on erinev, neil on erinev kõnetrakti kuju ja mõõtmed;
- ühe ja sama kõneleja hääldus on ajas muutuv ja sõltub palju kõneleja emotsionaalsest ja füüsilisest seisundist, kõnetempost, jne.

Kõnetuvastus sisaldab endas mitmeid erinevaid alamülesandeid, olulisemad neist on järgmised:

- häälikute tuvastamiseks vajalike tunnuste (10 kuni 20) leidmine signaalist kindla ajaintervalli (tavaliselt 10–20 msek) järel. Selleks kasutatakse spetsiifilisi signaalitöötlusalgoritme, erinevaid spektraalanalüüsi meetodeid ja kõnetaju mudeleid;
- tunnuste klassifitseerimine ja häälikukandidaatide leidmine. Põhiliselt kasutatakse Markovi varjatud mudeleid (HMM) ja neuronvõrke, mida eelnevalt on treenitud suure hulga kõnematerjali baasil;
- sõnahüpoteeside leidmine ja valik. Kasutatakse Markovi varjatud mudeleid (HMM) koos erinevate otsimisalgoritmidega (näiteks Viterbi algoritm).

Kõnetuvastussüsteemide iseloomustamiseks ja võrdlemiseks kasutatakse järgmisi parameetreid:

- kõnemaneer – isoleeritud sõnad või sidus kõne,
- kõnestiil – lugemine või spontaanne kõne,
- kasutajasõltuvus – kõnelejast sõltuv või sõltumatu,
- sõnastiku suurus – väike (< 100 sõna), keskmine (<10000), suur (< 100000 sõna),
- signaal/müra suhe – suur (> 30 dB) kuni väike (< 10 dB),
- erinevad tuvastuskorrektsuse näitajad jt.

Kõnetuvastussüsteeme on välja töötatud enamike suuremate keelte (inglise, prantsuse, saksa, hispaania, itaalia, hiina, jaapani jt.) jaoks, sõnastiku suurus ulatub juba üle 100000 sõna, tuvastuskorrektsus on paremate süsteemidel ca 95%.

3.2. Tarkvara eesti keele jaoks: mis on olemas

3.2.1. Lõppkasutajale mõeldud programmid

3.2.1.1. Kirjutaja abivahendid

Kirjutaja-abivahendeid on eesti keele jaoks loonud pms. Filosoft (poolitus, speller ja tesaurus); poolitusprogramme on teinud teisedki, nt. Indrek Hein ja Enn Saar. Lisaks kommertstoodetele on eesti keele jaoks olemas ka vabavarana levitatavale kontoritarkvarale OpenOffice sobiv speller ja poolitaja ning UNIXi keskkonnas vabavarana levitatav õigekirjakorrektor e. speller iSpell. Kõik Eestis kasutatavad tekstitöötlusprogrammid siiski eesti keele kontrollivahendite kasutamist ei võimalda (nt. Quark Xpress, StarOffice'i uuemad versioonid).

3.2.1.2. Dokumenditöötlus

Dokumenditöötluses kasutatakse Eestis keeletehnoloogiat vähem kui olemasolev tehnoloogiline baas seda võimaldaks. Üksikute eranditena võiks esile tuua Riigikantselei, kus sõnavormide muutlikkust arvestav dokumentide haldus- ja otsisüsteem on kasutusel alates 1996. aastast ja kus sellega on hõlmatud valitsuse otsused ja määrused, ning Eesti Õiguskeele Keskust, kus kasutatakse sõnavormide muutlikkust arvestavat terminite andmebaasi.

Eesti keelele on kohandatud optilise tekstituvastuse e. OCR tarkvara. Tegemist on Vene firma ABBYY (BIT Software), http://www.abbyy.ru programmi Fine Reader Pro täielikult eestindatud versiooniga, mis sai ajakirja „Arvutimaailm" preemia kui 1998. a. Eesti parim tarkvaratoode. Ta on mõeldud tekstide sisestamiseks arvutisse skänneri abil, võimaldades skänneriga sisestatud eestikeelset teksti ja tabeleid teisendada üldtunnustatud tekstiredaktorite ja tabelitöötlusprogrammide kujule. ABBYY poolt lisati programmile eesti keele tähtede ja sõnade identifitseerimise, mis kasutavad seejuures eesti keele spellerit.

3.2.1.3. Tõlkijate abivahendid ja lokaliseerimine

Tõlkijate abivahenditest kasutatakse Õiguskeele keskuses tõlkemälu (Trados), kus on oma sisemiseks kasutamiseks tehtud ka paralleelistaja e. joondaja. Nimelt on tõlkemälu loomiseks vaja originaal- ja tõlketeksti paralleelistamist e. joondamist, st. lähteteksti fragmentide ja nende tõlgete omavahelise vastavuse leidmist. Sellele alles järgneb lähteteksti fragmendile sobiva tõlke leidmine.

Eesti esimeseks lokaliseeritud ja eestindatud tarkvaravaldkonnaks oli raamatupidamis- jms. majandusarvestusega seotud tarkvara. Alates paketist Office XP on ka Microsofti poolt loodud kontoritarkvara (mis on omas klassis Eestis levinuim) järk-järgult eesti keelde tõlgitud. Ka Microsofti operatsioonisüsteem Windows XP on eestikeelne.

Vabatahtlike entusiastide poolt on eestindatud küllalt suur hulk vabavara. Eksisteerib mitu operatsioonisüsteemi Linux eestikeelset

varianti ning Linuxil töötav graafiline keskkond KDE on eestikeelne. Eestindatud kontoritarkvara pakett OpenOffice sisaldab ka eesti spellerit ja poolitajat. Eestikeelsed on brauserid Mozilla ja Opera, muusikamängimise programm Winamp, meiliprogramm Pegasus. Võru keelde on tõlgitud brauser Opera.

3.2.1.4. Masintõlge

Eesti pakub oma vene<->eesti masintõlkesüsteeme kaks Eesti firmat: Real Software ja EVE Systems. I. Hein EKIst on teinud sõnasõnalist tõlget tegeva brauseri.

3.2.1.5. Keeleõpe arvuti abil ja sõnaraamatud

Keeleõpe arvuti abil on lapsekingades: eksisteerib ainult üksikuid eesti keele õpetamiseks ja õppimiseks mõeldud programme: COMBO (Eve Systems), GuessWho (Anton Vylitok), LinguaMatch Pro (Nekstom) ning needki kasutavad pigem traditsioonilisi võtteid kui keeletehnoloogiat. Elektroonilise sõnastikke ja leksikone on seevastu mitmeid: inglise<->eesti sõnastikke pakuvad Festart, IBS ja I. Hein; vene<->eesti sõnastikke Real Software, Nekstom koos vene firmaga ABBYY; eesti<->mitmekeelseid sõnastikke pakuvad Soome firmad Sandstone ja Euroword Software; mitmesuguseid ükskeelseid sõnastikke pakuvad EKI, Filosoft, TEA ja mitmed üksikisikud internetis (vt. täpsemalt käesolevas ülevaates p. 2.3.1).

3.2.1.6. Kõnesüntees

On olemas eestikeelne kõnesüntesaator koos ekraanilugejaga, st. programm, mis arvuti ekraanil oleva eestikeelse teksti valjusti ette loeb.

3.2.1.7. Kõnetuvastus

Kõnetuvastuse alased tööd on algusjärgus, mistõttu laiatarbeprogramme veel ei ole.

3.2.1.8. Loomulikku keelt võimaldavad kasutajaliidesed

Loomulikku keelt võimaldavate kasutajaliideste osas tehakse Eestis uurimistööd dialoogi modeleerimise ja dialoogiaktide äratundmise vallas.

3.2.2. Teadus- ja arendustöödeks mõeldud programmid

3.2.2.1. Sõnavormide analüüs ja süntees

Sõnavormide analüüsi tegevaid programme nimetatakse morfoloogilisteks analüsaatoriteks, sünteesi tegevaid programme morfoloogilisteks süntesaatoriteks. Kuigi eesti keele morfoloogia on teadagi keeruline, on loodud mitmeid morfoloogilisi analüsaatoreid ja süntesaatoreid. Praegu on teoreetilistel ja praktilistel eesmärkidel kasutatavad EKI (http://www.eki.ee) ja OÜ Filosofti (http://www.filosoft.ee) analüsaatorid ja süntesaatorid. Filosofti analüsaator on näiteks aluseks mitmele eesti keele spellerile ja poolitajale, samuti Riigikantselei täistekstiandmebaasis TRIP alates 1996. aastast kasutatud sõnade algvormide leidmisele ning on üheks komponendiks difoonidel põhinevas eesti kõne süntesaatoris. EKI analüsaatorit kasutatakse EKIs sõnastikukirjete genereerimisel jm. Nii EKI kui Filosofti programmid võimaldavad analüüsida ja sünteesida nii sõnu, mis on nende sõnastikes, kui sisaldavad ka oletajat nende sõnade analüüsi ja sünteesi jaoks, mida sõnastikes pole.

TÜs tegeldakse ka 2-tasemelise morfoloogilise analüsaatori ja süntesaatori väljatöötamisega (Roosmaa jt. 2003). Tema erinevus olemasolevatest seisneb sõnastiku ja morfoloogiareeglite esitusviisis. 2-tasemeline morfoloogiamudel on praegu maailmas levinuim mudel morfoloogiaprogrammide tegemiseks.

3.2.2.2. Grammatiline analüüs ja süntees

Grammatilist analüüsi tegevaid programme nimetatakse süntaktilisteks analüsaatoriteks, sünteesi tegevaid programme süntaktilisteks süntesaatoriteks. Viimaseid eesti keele jaoks olemas ei ole.

Esimene ja lihtsam etapp lause grammatilisel analüüsil on mitmeti tõlgendatavate morfoloogiliste vormide tõlgenduste hulgast selliste variantide valimine, mis just antud konteksti sobivad. Seda nimetatakse morfoloogiliseks ühestamiseks. Eesti keele jaoks on olemas kaks morfoloogilist ühestajat, mis mõlemad on kasutatavad nii teoreetilistel kui praktilistel eesmärkidel. Tartu Ülikoolis on loodud kitsenduste grammatika formalismil põhinev, reeglipõhine ühestaja; OÜ Filosoftis on loodud statistiline, Markovi varjatud mudelil põhinev ühestaja. TÜ ühestajat on kasutatud sisukokkuvõtete automaatseks tegemiseks ja nimisõnafraaside automaatseks tuvastamiseks. Filosofti ühestajat on kasutatud sagedussõnastiku (Kaalep, Muischnek 2002) tegemiseks ja difoonidel põhinevas eesti kõne süntesaatoris.

Eesti keele jaoks on olemas üks süntaktiline analüsaator. See on loodud TÜs ja põhineb kitsenduste grammatika formalismil. Teda on kasutatud sisukokkuvõtete automaatseks tegemiseks ja nimisõnafraaside automaatseks tuvastamiseks.

3.2.2.3. Semantiline analüüs ja süntees

Eesti keele semantilise analüüsi ega sünteesiga seni otseselt tegeldud pole. Küll on tehtud ettevalmistavaid töid. TÜs on loomisel WordNeti tüüpi eesti keele tesaurus (Orav, Vider 2002; Kahusk, Vider 2002), milles seisuga oktoober 2003 on 11,5 tuhat sünohulka. TÜs on loodud ka katseline semantilise ühestamise programm (Vider, Kaljurand 2001; Kahusk, Orav, Õim 2001; Kahusk 2002; Kahusk, Kaljurand 2002), mis tekstis esinevatele mitmetähenduslikele sõnadele leiab konteksti sobiva tähenduse. Mitmetähenduslikuks peetakse seejuures sõnu, millel on tesauruses mitu tähendust (Kahusk, Vider 2002).

3.2.2.4. Pragmaatiline analüüs ja süntees

Pragmaatikaga tegeldakse TÜs suulise kõne ja dialoogi uurimise kontekstis (Koit 2003a, 2003b; Hennoste jt. 2003).

3.2.2.5. Kõnesüntees

Difoonidel põhinev eestikeelne kõnesüntesaator on sellisel küpsusastmel, et ta on kasutusel praktilistel eesmärkidel, nt. nägemis- ja kõnepuudega inimeste poolt. Samas jätkub uurimistöö sünteeskõne loomulikkuse tõstmiseks. Kõnekvaliteedi seisukohalt mängib erinevatest kõnesünteesi moodulitest – difoonide andmebaas, prosoodiamudel, teksti lingvistiline töötlus – olulisimat rolli prosoodiamudel, selle edasiarendusele on fokuseeritud lähiaastate uurimistöö. Sünteeskõne kvaliteedi hindamiseks kasutatakse rahvusvaheliselt aktsepteeritud metoodikat.

Artikulatoorse ega formantsünteesiga Eestis praegu ei tegelda.

3.2.2.6. Kõnetuvastus

On tehtud üksikuid esialgseid katsetusi piiratud sõnavara tuvastamisel. On välja töötatud piiratud sõnastikuga (50 sõna) tuvastussüsteemi prototüüp (Meister 2001; Meister jt. 2001), numbrituvastuse prototüüp (Alumäe 2001; Alumäe jt. 2003) ja teostatud eksperimente sidusa kõne tuvastamiseks.

Antud valdkonnaga on seotud ka kõnelejatuvastus. On uuritud neuronvõrkude rakendamist kõnelejatuvastusülesannete lahendamiseks (Altosaar, Meister 1995; Meister 1998), käimas on koostöö Helsingi Ülikooli foneetikaosakonnagakõnelejaspetsiifiliste tunnuste analüüsi alal (Iivonen jt. 2001; Meister 2002).

3.3. Tarkvara eesti keele jaoks: mida oleks vaja

Kui lühidalt kokku võtta, siis kõige efektiivsem tee eesti keele spetsiifilise tarkvara loomiseks on see, et kohandada maailmas olemas olev tehnoloogia eesti keelele sobivaks. Sellisele kohandamisele allub tehnoloogia, mis juba loomise hetkel oli keelest sõltumatu, s.t. keeleressursse eksplitsiitselt kasutav tehnoloogia. Asendades ühe keele ressursi (korpuse või sõnastiku) teise keele omaga, saamegi luua teisele keelele omast tarkvara.

See ei tähenda muidugi, et konkreetse tehnoloogia ülekandmine oleks rutiinne ja lihtne töö. Näiteks selleks, et kasutada eesti keele morfoloogiliseks ühestamiseks paljude muude keelte peal realiseeritud tehnoloogiaid, nii statistilisi kui reeglipõhiseid, tuli lisaks morfoloogiliselt märgendatud korpuse tegemisele vaadata kriitiliselt üle eesti keele morfosüntaktiliste kategooriate süsteem ja paljude konkreetsete sõnade puhul nende sõnaliigilist kuuluvust täpsustada. Samuti tuli olemasolevast tehnoloogiast nii põhjalikult aru saada, et mõista, kas üleskerkinud probleemid on tingitud tehnoloogia piiratusest (ükski tehnoloogia pole kõikvõimas), tema valest rakendamisest, eesti keele omapärast või hoopis meie teadmiste piiratusest eesti keele kohta.

Alljärgnevalt kirjeldame lisaks konkreetsele vajalikule tarkvarale ka selle loomiseks vajalikke keeleressursse.

3.3.1. Lõppkasutajale mõeldud tarkvara

3.3.1.1. Kirjutaja abivahendid

Esmajärjekorras oleks vaja grammatika kontrollijat ja teksti mõistetavust hindavat programmi. Ülesannet lihtsustab see, et sellised põhiabivahendid nagu õigekirja kontrollija ja poolitaja on olemas. Samuti on olemas piirangute grammatikal põhinev süntaksi analüsaator.

Eeldab:

1) “tüüpiliste vigade” korpuse loomist, kusjuures vead sõltuvad kirjutaja emakeelest, tekstitüübist jne;

2) süntaktilise analüüsi probleemide lahendamist just korrektsuse ja normatiivsuse vaatevinklist, kuid seda saab teha järkjärgult: teatud tüüpi fraaside sees tehakse teatud tüüpi vigu: nimisõnafraasides ühildumisvigu, verbifraasides rektsioonivigu jne.

Probleemiks võib olla eesti keele süntaksi ikka veel mittepiisav teoreetiline läbiuuritus, eriti just sellisest formaalsest vaatepunktist nagu on vaja keeletehnoloogiliseks arendustööks.

3.3.1.2. Dokumenditöötlus

Infootsing täistekstidest märksõnade (mõistete) järgi, kus kasutaja ei pea mõtlema eesti keele morfoloogiliste iseärasuste peale: nt. et „suveaja" saaks üles leida, kui otsida sõna „suveaeg".

Eeldab:

1) märksõnade (tüüpiliselt terminid, ka mitmesõnalised) tuvastamist tekstis; kuna valdav osa termineid on nimisõnafraasid, siis on see vaadeldav nimisõnafraaside tuvastamise allülesandena, mis omakorda eeldab süntaktilise analüsaatori kasutamist;
2) kõrgekvaliteedilist morfoloogilist oletajat ja ühestajat, sest enamasti pakuvad infootsingul huvi just sellised sõnad ja fraasid, mida sõnastikes ei ole: pärisnimed (nt. isikud, tooted, firmad, institutsioonid), teoste pealkirjad, uhiuued terminid.

Abiks oleks ka vastava ainevaldkonna terminite tesaurus, mis võimaldab otsingut korraldada lisaks vahetult antud terminite ka nendega relevantsetes semantilistes seostes olevate terminite ja muude väljendite kaudu.

Mitmekeelne infootsing: vajalikku infot otsitakse tekstidest, mis lisaks eesti keelele võivad olla ka teistes keeltes (eriti aktuaalne EL raames).

Eeldab:

1) sama, mida infootsing;
2) mitmekeelseid sõnastikke.

Abiks oleks ka vastavate keelte vastava ainevaldkonna materjali sisaldav mitmekeelne tesaurus, kus mõisted on keeliti seotud.

Dokumentide liigitamine ja refereerimine

Eeldab:

1) sama, mida infootsing;
2) vastava ainevaldkonna mõistelist liigendust (nt. tesauruse üldkategooriate tasemel) ja lisaks morfoloogilisele analüüsile süntaktilist analüüsi lause ulatuses, mis esialgu võib olla suhteliselt lihtsustatud variant.

3.3.1.3. Tõlkijate abivahendid ja masintõlge

Eeldavad:

1) paralleelkorpusi,
2) morfoloogilist analüüsi ja sünteesi,
3) süntaktilist analüüsi ja sünteesi,
4) mitmekeelseid sõnastikke.

Ükskõik, kas masintõlke ja tõlkija abivahendite aluseks võetakse transfer-meetod, tõlkemälu-meetod või müranivooga kanali meetod, kvaliteetse masintõlkeni jõudmine võtab igal juhul veel palju aega. Praktiliselt kasulike tõlkija-abivahendite loomine on oletatavasti saavutatav lähemas perspektiivis kui masintõlge. Samas on paljud masintõlkeks vajalikud ressursid ja tarkvara-komponendid kasutatavad ka muudel aladel kui kitsalt masintõlkes ja tõlkija abivahendites: nt. info-otsingus, grammatikakontrollijas, keeleõppes.

3.3.1.4. Keeleõpe arvuti abil ja sõnaraamatud

Ehkki keeleõpe arvuti abil ei puuduta otseselt eesti keeletehnoloogiat, aitab ta siiski laiendada eesti keele (elektroonilist) kasutussfääri ja seega toetab ka keeletehnoloogiat.

Oluline oleks keeletehnoloogiat kasutavate sõnaraamatute loomine (nt. et algvorme ja mitmesõnalisi fraase tekstist automaatselt leida). Sellised sõnaraamatud oleksid ka samm tõlkijate abivahendi-

te ning masintõlke suunas. Ka kõnesignaalide analüüsivahendeid saaks kasutada eesti keele häälduse õpetamisel vene vm. võõrkeelt kõnelevatele inimestele.

Eeldused:

1) sama, mis tõlkijate abivahenditel ja masintõlkel;
2) kõneanalüüsi vahendid.

3.3.1.5. Loomulikku keelt võimaldavad kasutajaliidesed

Automaatsete eestikeelsete infosüsteemide väljatöötamine eeldab:

1) eestikeelse dialoogi struktuuri uuringuid kindlalt piiritletud valdkondades;
2) eestikeelse kõnetuvastuse väljatöötamist.

3.3.1.6. Kõnesüntees

Juba praegu oleks võimalik luua mitmesuguseid kõnesünteesikasutatavaid rakendusi:

- pimedate lugemismasin. See oleks kõnesüntesaator koos skänneri ja optilise tekstituvastuse programmiga, mis võimaldab trükitud eestikeelset teksti – raamatuid, ajalehti jms. – pimedatele ette lugeda,
- e-posti ettelugemine telefoni teel. Helistades e-posti serverisse, oleks võimalik kuulata kõnesüntesaatori abil loetud e-kirja.

3.3.1.7. Kõnetuvastus

Eeldab kõnetuvastuse alaseid baasuuringuid, seetõttu pole lähiaastatel vastavaid rakendusi oodata. Kaugemas perspektiivis on võimalikud kõnetuvastuse rakendused järgmised:

- teksti dikteerimine arvutile,
- seadmete-arvutite hääljuhtimine,
- inimene-masin dialoogsüsteemid jm.

3.3.2. Teadus- ja arendustöödeks mõeldud programmid

3.3.2.1. Sõnavormide analüüs ja süntees

Olemasolev eesti morfoloogia-alane tarkvara on mõeldud tänapäeva kirjakeele töötlemiseks. Vaja oleks hõlmata ka muud keelesfäärid:

1) kõnekeel ja suuline kõne (ei saa toetuda suure ja väikese tähe eristamisele; peab hakkama saama paljude katkendlike ja vigaste sõnadega),
2) murdekeel,
3) vana kirjakeel.

Eeldus:

1) suulise keele korpus,
2) kõnekorpus,
3) murdekorpus,
4) vana kirjakeele korpus.

Tekstis esinevate mitte-sõnade (lühendite, arvude, valemite), üliharuldaste sõnade (erialaterminite, pärisnimede) ja kirjavigadega sõnade analüüsiks ei sobi samad reeglid, mis keskse sõnavara analüüsiks. Samas kuuluvad ka need morfoloogia mõttes mitte-standardsed elemendid eesti keelde ja nõuavad tõlgendamist. Keele perifeerias toimuv ennustab sageli seda, mis keele keskses osas veidi aja pärast toimuma hakkab.

Eeldus:

1) suur tekstikorpus (et mittestandardsed nähtused esile tuleks).

3.3.2.2. Grammatiline analüüs ja süntees

Vaja oleks:

1) minna olemasolevast eesti keele süntaksi analüsaatorist, mis piirdub pindanalüüsiga, edasi ja luua lause süntaktilist struktuuri täpsemalt kirjeldav analüsaator,
2) luua süntaktiline süntesaator,

3) parandada olemasolevate morfoloogiliste ühestajate kvaliteeti,
4) luua morfoloogilised ühestajad ka muude keelesfääride kui kirjakeele jaoks, eelkõige kõne jaoks.

Eeldus:
1) süntaktiliselt märgendatud korpus,
2) morfoloogiliselt märgendatud korpus,
3) formaliseeritud grammatikakirjeldus.

3.3.2.3. Semantiline analüüs ja süntees

Vaja oleks:
1) suurt leksikaalsemantilist andmebaasi,
2) semantilise ühestamise tarkvara, st. programmi, mis teksti igale sõnale oskaks juurde märkida, millist konkreetset tähendust selles suures leksikaalsemantilises andmebaasis antud kontekstis tuleks kasutada.

3.3.2.4. Pragmaatika

Eeldab teoreetilisi uuringuid eestikeelsete tekstide ja diskursuste ülesehituse kohta.

Vaja oleks:
1) teksti sidususe suhtes märgendatud tekstikorpust (kus oleks nt. märgitud, millisele tekstiosale viitab asesõna),
2) kõneaktide suhtes märgendatud dialoogikorpust.

3.3.2.5. Kõnesüntees

Sünteeskõne kvaliteedi tõstmiseks on vajalik eelkõige kõne prosoodilise struktuuri (meloodia) täiuslikum modelleerimine.

Eeldab:
1) lause automaatse süntaktilise analüüsi lahendamist;
2) inimkõne prosoodilise struktuuri põhjalikku analüüsi,

3) tekstis esinevate mitte-sõnade (numbrid, lühendid, valemid jne.) analüüsi ja vastavate sõnaliste väljendite morfoloogilist sünteesi.

3.3.2.6. Kõnetuvastus

Kõnetuvastuseks vajalik tehnoloogia on maailmas põhimõtteliselt olemas ja seda rakendatakse edukalt põhiliselt mitte-aglutineerivate keeletüüpide korral. Aglutineerivate-flekteerivate keelte, sh. eesti keele puhul vajab eelnimetatud tehnoloogia olulisi keele-spetsiifilisi ja põhimõttelisi täiendusi.

Eestikeelse kõnetuvastuse väljatöötamine eeldab:

1) baastehnoloogia soetamist,
2) uuringuid ja eksperimente tuvastuseks sobivate kõnesegmentide (foneemid, difoonid, trifoonid, silbid) väljaselgitamiseks,
3) mahuka kõne andmebaasi olemasolu nii uuringuteks kui ka süsteemi treenimiseks ja testimiseks.

KOKKUVÕTE

Käesoleva lisa ülesandeks oli esitada ülevaade eesti keelt puudutavate keeletehnoloogiaalaste tööde seisust ja sellest lähtuvalt sõnastada argumenteeritud ülesanded eesti keele keeletehnoloogilise toe loomiseks aastaks 2010. Keele tehnoloogilise toe määratluse järgi hõlmab see elektroonilisi keeleressursse, keeletöötlustarkvara ja keeletehnoloogilisi rakendussüsteeme.

Ülevaade näitab, et eesti keele osas on tegeldud kõigi kolme valdkonnaga, aga erineval määral.

Enim on edu saavutatud keeleressursside arendamisel. Keeleressursid on elektroonilised teksti- ja kõnekorpused ja kõneandmebaasid ning arvutileksikonid. Keeleressursid on aluseks keeletarkvara väljatöötamisele ja on loomulik, et eesti keele tehnoloogilise toe loomine on alanud sellest valdkonnast.

Eesti keele arvutiressursid on teiste Euroopa keeltega võrreldes rahuldaval tasemel ja kui töid saab jätkata planeeritud viisil, siis on aastaks 2010 kavandatud tase saavutatav.

Kirjutatud keele korpusi on ette nähtud lähiaastatel koguda 100 miljonit sõna ja see saavutatakse lähema 3 aasta jooksul. Aastaks 2010 kogutakse 200 miljoni sõnaline korpus. Raskusi on siiski eestikeelse ilukirjanduse ja eestikeelsete teadustekstide kogumisega.

Suulise kõne korpust on kogutud 1996. aastast. 2003.aasta septembrikuu seisuga on korpuses u. 600 000 tekstisõna. Eesmärk – koguda 2010. aastaks 3 miljonit tekstisõna – on saavutatav. Suulise kõne korpus sisaldab ka dialoogikorpust, mis on aluseks telefonipõhiste suhtlussüsteemide väljatöötamisele. 2003. aastal alustati ühtlasi kõneandmebaaside loomist (vähemalt 2000 erinevat kõnelejat), mis on eelduseks kõnetuvastussüsteemi väljaarendamisele.

Keeleressursside teine oluline alaliik on arvutileksikonid: üksja mitmekeelsed sõnastikud, mis on vajalikud keeletehnoloogilistes rakendussüsteemides. Eesti keel on arvutileksikonidega hästi esin-

datud, ehkki need ei ole loodud alati keeletehnoloogilisi rakendusi silmas pidades. Arvutileksikonide osa (vt. osa 2. Arvutileksikonid) esitab detailse ülevaate arvuti abil töödeldavatest eesti keele sõnastikest. Ülevaade näitab, et olulisemad tööd antud valdkonnas aastani 2010:

- ühendada erinevad üks- ja mitmekeelsed sõnastikud üheks andmebaasiks, kus iga sõnakirje sisaldab morfoloogilist ja süntaktilist informatsiooni, mis on vajalik automaatses tekstitöötluses;
- täiendada olemasolevat eesti keele semantilist andmebaasi (eesti wordnetti) kuni 100 000 sõnani, et seda saaks kasutada üldkeelele orienteeritud info-otsi- ja masintõlkesüsteemides;
- luua (mitmekeelsed) terminiandmebaasid (seejuures kasutades üldkeele – wordneti põhimõtteid) erialade jaoks, mis on esmajoones tähtsad keeletehnoloogilise toe arendamisel.

Ülevaate kolmas osa kirjeldab eesti keele töötlemise tarkvara ja selle rakendusi, taustaks on esitatud info olukorrast selles valdkonnas maailmas. Ülevaatest nähtub, et eesti keele töötlemise tarkvara on võrdlemisi ebaühtlases seisus. On aga fikseeritud suunad, milles esmajärjekorras on vaja edasi töötada. Lähtealused selleks on olemas tänu sellele, et eesti keele morfoloogilise analüüsi/sünteesi programmid on välja töötatud, nagu ka süntaksi pindanalüüsi programm ja tekst-kõne sünteesiprogramm. Morfoloogiaanalüsaatori põhjal on loodud õigekirjakorrektor ja poolitaja. Puudu on süntaksi süvaanalüüsi programm ja (lausete/teksti) semantilise analüüsi programm.

Käsitletaval perioodil tuleb välja töötada järgmised programmid (lisaks olemasolevate programmide täiustamisele):

- automaatne kõnetuvastus;
- grammatikakorrektor;
- tõlkeabi- ja masintõlkeprogrammid;
- sisukokkuvõtete tegemise programmid;
- eestikeelset infodialoogi modelleeriv programm;
- sisupõhised infootsiprogrammid.

Analüüs näitab, et Eestis on olemas spetsialistid keeleressursside, keeletarkvara ja rakendussüsteemide väljatöötamiseks, aga ka võimalused täiendavate spetsialistide koolitamiseks Tartu Ülikooli arvutilingvistika ja keeletehnoloogia erialal.

KASUTATUD KIRJANDUS

Alumäe, T. 2001. Eestikeelse kõne tuvastus: prototüübi loomine. Tallinna Tehnikaülikool. Tallinn [Magistritöö].

Alumäe, T., Võhandu, L. 2003. Piiratud ulatusega eestikeelne kõnetuvastus. – Eesti Keele Instituudi toimetised 12. Tallinn.

Altosaar, T., Meister, E. 1995. Speaker recognition experiments in Estonian using multi-layer feed-forward neural nets. – Proceedings of Eurospeech'95. Vol.1. Madrid, 333–337.

Altosaar, T., Karjalainen, M., Vainio, M., Meister, E. 1998. Finnish and Estonian speech applications developed on an object-oriented speech processing and database system. – Workshop on Speech Database Development for Central and Eastern European Languages, Granada, Spain, May .

Automatic Morphology of Estonian 1. (Research Reports). 1994. Toim. Viks, Ü. Tallinn: Eesti Keele Instituut.

Automatic Morphology of Estonian 2. (Research Reports). 1995. Toim. Viks, Ü. Tallinn: Eesti Keele Instituut.

Current Issues in Computational Linguistics: In Honour of Don Walker. 1997. Ed. by Zampolli, A., Calzolari, N., Palmer, M. Kluwer Academic Publishers.

Danzin, A. and the Planning Study Group 1992. Towards a European Language Infrastructure. Report to the Commission of European Communities. 31.March.

Eesti keele formaalne grammatika. 2001. Koost. Roosmaa, T., Koit, M., Muischnek, K., Müürisep, K., Puolakainen, T., Uibo, H. Tartu Ülikooli arvutiteaduse instituut. Tartu: Tartu Ülikooli Kirjastus.

Eesti kirjakeele sagedussõnastik. 2002. Koost. Kaalep, H.-J., Muischnek, K. Tartu: Tartu Ülikooli kirjastus.

Eilsen, K. 2000. Georg Mülleri sõnastik arvutis. – Pipliakielestä kirjakieleksi. (Kotimaisten kielten tutkimuskeskuksen julkaisuja 105.) Helsinki: Kotimaisten kielten tutkimuskeskus, 319–327.

Ehasalu, E., Habicht, K., Kingisepp, V-L., Peebo, J. 1997. Eesti keele vanimad tekstid ja sõnastik. Tartu Ülikooli eesti keele õppetooli toimetised 6. Tartu.

Fellbaum, C. 1998. Introduction. – WordNet: An Electronic Lexical Database. Ed. by Fellbaum, C. Cambridge, Massachusetts: MIT Press, 1–19.

Fillmore, C. J. 1977. Scenes-and-frames semantics, Linguistic Structures Processing. – Fundamental Studies in Computer Science, No. 59. Ed. by Zampolli, A. North Holland Publishing.

Fillmore, C.J., Baker, C. F., Lowe, J.B. 1997. A frame-semantic approach to semantic annotation. – Proceedings of the SIGLEX workshop "Tagging Text with Lexical Semantics: Why, What, and How?" [WWW]

- http://www.icsi.berkeley.edu/~framenet/

Habicht, K., Kingisepp, V-L., Pirso, U., Prillop, K. 2000 Georg Mülleri jutluste sõnastik. Tartu Ülikooli eesti keele õppetooli toimetised 12. Tartu.

Heid, U., Krüger K. 1996. A multilingual lexicon based on Frame Semantics. – Proceeding of AISB96 Workshop on Multilinguality in the Lexicon. Ed. by Cahill, L. and Roger, E. University of Sussex, UK.

Hennoste, T., Muischnek, K., Potter, H., Roosmaa, T. 1993. Tartu Ülikooli kirjakeele korpus: ülevaade tehtust ja probleemidest. – Keel ja Kirjandus, 10, 587–600.

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. Ed. by Õim, H. 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. 2000. Eesti suulise kõne uurimine: transkriptsioon, taust ja korpus. – Keel ja Kirjandus 2, 91–106.

Hennoste, T., Muischnek, K. 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 183 – 218.

Hennoste, T., Lindström, L., Rääbis, A., Toomet, P., Vellerind, R. 2000. Eesti suulise kõne korpus ja mõnede allkeelte võrdluse katse. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeelteaduse õppetooli toimetised 1. Toim. Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 245–283.

Hennoste, T., Kaalep, H.-J., Muischnek, K., Paldre, L., Vaino, T. 2001. The Tartu University Corpus of Estonian Literary Language. – Congressus Nonus Fenno-Ugristarum Pars V. Tartu, 337–344.

Hennoste, T., Lindström, L., Rääbis, A., Toomet, P., Vellerind, R. 2001. Tartu University Corpus of Spoken Estonian. – Congressus Nonus Fenno-Ugristarum Pars V. Tartu, 345–351.

Hennoste, T., Koit, M., Kullasaar, M., Rääbis, A., Vutt, E. 2002. Eesti dialoogikorpuse loomise probleemid. – Tähendusepüüdja. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim. Pajusalu, R. ja Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 143–160.

Iivonen, A., Harinen, K., Keinänen, L. Liisanantti, H., Meister, E., Tuuri, L. 2001. Moniparametrinen puhujantunnistus. 21. Fonetiikan Päivät, Turku 4.–5.1.2001. – Publications of the Department of Finnish and General Linguistics of the University of Turku. Ed. by Ojala, S., Tuomainen, J. Turku, 81–95.

Kaalep, H.-J. 1999. Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös. Dissertationes philologiae estonicae Universitatis Tartuensis 7. Tartu Ülikool. Tartu. [Doktoritöö].

Kaalep, H.-J., Muischnek, K., Müürisep, K., Rääbis, A., Habicht, K. 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest. – Keel ja Kirjandus 9, 623–633.

Kaalep, H.-J., Vaino, T. 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 87 – 99.

Kaalep, H.-J., Vaino, T. 2001. Complete Morphological Analysis in the Linguist's Toolbox. – Congressus Nonus Internationalis Fenno-Ugristarum Pars V. Tartu, 9 – 16.

Kaalep, H-J., Muischnek, K. 2002a. Using the Text Corpus to Create a Comprehensive List of Phrasal Verbs. – Proceedings LREC 2002. Third International Conference on Language Resources and Evaluation.Vol. 1. Ed. by Rodríguez, M. G., Suarez Araujo, C. P. Granada, 101–105.

Kaalep, H-J., Muischnek, K. 2002b. Püsiühendite leidmine teksti abil. – Tähendusepüüdja. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim. Pajusalu, R. ja Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 172–184.

Kahusk, N., Orav, H., Õim H. 2001. Sensiting inflectionality: Estonian task for SENSEVAL 2. – SENSEVAL–2 Workshop Proceedings. 25–28.

Kahusk, N. 2002. A Lexicographer's Tool for Word Sense Tagging According to WordNet. – Workshop on Wordnet Structures and Standardisation, and How these Affect Wordnet Applications and Evaluation. LREC 2002 Workshop Proceedings. 1–7.

Kahusk, N. and Vider, K. 2002. Estonian Wordnet benefits from word sense disambiguation. – Proceedings of the 1st International Global WordNet Conference. Central Institute of Indian Languages. Mysore, ndia, 26–31.

Kahusk, N., Kaljurand, K. 2002. Semyhe tulemusi: kas tasub naise pärast WordNet ümber teha? – Tähendusepüüdja. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim. Pajusalu, R. ja Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 185–195.

Koit, M. 2002a. Märgendatud dialoogikorpus: miks ja kuidas? – Konverents „Rakenduslingvistika Eestis". Teesid. Tallinn, 31–32.

Koit, M. 2002b. Kommunikativnye strategii v informacionno-spravochnom dialoge (na materiale estonskogo korpusa dialogov). – Proceedings DIALOG–2002.Vol. 2. Moskva: Nauka, 283–290.

Koit, M. 2003. Märgendatud dialoogikorpus kui keeleressurss. – Toimiv Keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 12. Toim. Langemets, M., Sahkai, H., Sepper, M.-M. Tallinn: Eesti Keele Sihtasutus, 119–136.

Kuusik, E. 1996. Eesti tüvemuutuste süsteemi modelleerimine. Eesti Keele Instituut. Tallinn. [Magistritöö].

Langemets, M. 2000. Sõnaraamatu arvutilingvistiline analüüs. Eesti Keele Instituut. Tallinn. [Magistritöö].

Langemets, M. 2002. Eesti Keele Instituudi elektrooniline keelevara. – A&A, 5.

Lindström, L. 2001. Eesti murrete korpuse iseloomustus argivestlustega võrrelduna. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks 12. märtsil 2001. Tartu Ülikooli eesti keele õppetooli toimetised 17. Koost. ja toim. Kasik, R. Tartu: Tartu Ülikooli Kirjastus, 212–221.

Lindström, L., Lonn, V., Mets, M., Pajusalu, K., Teras, P., Veismann, A., Velsker, E., Viikberg, J. 2001. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks 12. märtsil 2001. TÜ eesti keele õppetooli toimetised 17. Koost. ja toim. Kasik, R. Tartu: Tartu Ülikooli Kirjastus, 186–211 .

Meister, E., Eek, A. 1996. Estonian Phonetic Database: Development and Realisation. – Proceedings of the Second International Baltic Workshop on Databases and Information Systems. Vol. 2. Tallinn, 159–168.

Meister, E. 1998. Kõnelejatuvastuse eksperimendid neuronvõrkudel. Tallinn. [Magistritöö].

Meister, E., Eek, A., Altosaar, T., Vainio, M. 1999. The Estonian Phonetic Database in the Quicksig Object-Oriented Environment. – Proceedings of the International Workshop on Computational Linguistics and its Applications. Vol. 2. Tarusa, 347–350.

Meister, E., Eek, A., Altosaar, T., Vainio, M. 2000. Object-Oriented Access to the Estonian Phonetic Database. – Proceedings of the Second International

Conference on Language resources and Evaluation. Vol.1. Athens, Greece, 269–272.

Meister, E. 2001. Towards speech recognition in Estonian. 21. Fonetiikan Päivät, Turku 4.–5.1.2001. – Publications of the Department of Finnish and General Linguistics of the University of Turku. Ed. by Ojala, S., Tuomainen, J. Turku, 59–70.

Meister, E., Lobanov, B., Vahisalu, R., Levkovskaya, T., Kisialou, V., Tatter, P., Lasn, J. 2001. Spoken Dialogue System for Mobile Parking. – Proceedings of the International Workshop SPEECH and COMPUTER (SPECOM'2001). Moscow, Russia, 123–126.

Meister, E. 2002. Kõneleja-spetsiifiliste tunnuste otsingul. – Tähendusepüüdja. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim. Pajusalu, R. ja Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 266–284.

Meister, E., Lasn, J., Meister, L. 2002. Estonian SpeechDat: a project in progress. – The Phonetics Symposium. Ed. by Korhonen, P. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing. Espoo, 21–26.

Meister, E., Lasn, J., Meister, L. 2003. Development of the Estonian SpeechDat-like Database. – Proceedings of Eurospeech'2003. Vol. 2. Geneva, 1601–1604.

Mihkla, M., Eek, A., Meister, E. 1998a. Creation of the Estonian Diphone Database for Text-to-Speech Synthesis. – Proceedings of the Finnic Phonetics Symposium, Pärnu. Linguistica Uralica 3, 334–340.

Mihkla, M., Eek, A., Meister, E. 1998b. Text-to-speech Synthesis of Estonian. – Proceedings of 6th European Conference on Speech Communication and Technology. Vol. 5. Budapest, 2095–2098.

Mihkla, M., Eek, A., Meister, E. 1999. Diphone Synthesis of Estonian. – Proceedings of the International Workshop on Computational Linguistics and its Applications. Vol. 2. Tarusa, 351–353.

Mihkla, M., Meister, E., Eek, A. 2000. Eesti keele tekst-kõne süntees: grafeem-foneem teisendus ja prosoodia modelleerimine. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 309–319.

Mihkla, M., Meister, E., Eek, A., Hein, I., Tatter, P. 2000. Non-words interpreter, prosody generator and screen reader for the Estonian text-to-speech synthesizer. – Proceedings of the International Workshop Dialogue'2000. Computational Linguistics and Its Applications. Vol. 2 (Applications). Ed. by Narin'yani, A. S. Protvino, 399–407.

Mihkla, M., Meister, E., Eek, A., Lasn, J. 2001. Testing the quality of Estonian text-to-speech synthesis. 21. Fonetiikan Päivät, Turku 4.–5.1.2001.

– Publications of the Department of Finnish and General Linguistics of the University of Turku. Ed. by Ojala, S., Tuomainen, J. Turku, 40–45.
Mihkla, M., Meister, E., Kiissel, I., Lasn, J. 2001. Evaluation the quality of Estonian text-to-speech synthesis and diphone corrector for the TTS system. – Proceedings of the International Workshop Dialogue'2001. Vol. 2 (Applications). Aksakovo, 385–390.
Mihkla, M., Meister, E., Lasn, J. 2001. Quality Evaluation of Estonian Text-to-Speech Synthesis. – Proceedings of the International Workshop SPEECH and COMPUTER (SPECOM'2001). Moscow, Russia, 163–166.
Mihkla, M., Meister, E. 2002. Eesti keele tekst-kõnesüntees. – Keel ja Kirjandus 2, 88–97; 3, 173–182.
Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. 1993. Introduction to WordNet: an on-line lexical database. [WWW] – ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps
Müürisep, K. 1998a. Eesti keele kitsenduste grammatika süntaksianalüsaatorist. – Keel ja Kirjandus 1, 47–56.
Müürisep, K. 1998b. Syntactic analysis of Estonian using Constraint Grammar. – Proceedings of International Workshop Dialogue'98: Computational Linguistics and its Applications. Kazan, 619–625.
Müürisep, K. 1999. Determination of Syntactic Functions in Estonian Constraint Grammar. – Proceedings of EACL'99. Bergen, 291–292.
Müürisep, K. 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu Ülikool. Tartu [Doktoritöö].
Orav, H. 1998. Eesti keele direktiivverbide semantilise välja struktuur tesaurusena. Tartu Ülikool. Tartu. [Magistritöö].
Orav, H., Vider, K. 2002. Kas tesaurus ja tekstid lähevad kasutuses kokku? – Tähendusepüüdja. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Toim. Pajusalu, R. ja Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 297–303.
Orav, H., Muischnek, K. 2002a. Eesti keele korpused ja arvutileksikonid – mis on olemas ja mida veel vaja on. – Arvutimaailm 8, 15–18.
Orav, H., Muischnek, K. 2002b. Keeleressursid – mis ja milleks. – A&A 5, 47–52. Puolakainen, T. 2002. Eesti keele arvutigrammatika: morfoloogiline ühestamine. Dissertationes Mathematicae Universitatis Tartuensis. Tartu Ülikool. Tartu [Doktoritöö].
Roosmaa, T., Koit, M., Muischnek, K., Müürisep, K., Puolakainen, T., Uibo, H. 2003. Eesti keele arvutigrammatika: mis on tehtud ja kuidas edasi? – Keel ja Kirjandus 3, 192–209.
Saluveer, M., Õim, H. 1985. Frames in linguistic description. – Quaderni di Semantica. Vol. 6, no. 2, 282–292.

Tavast, A. 2002. Onomasioloogia ja semasioloogia vahekorrast oskuskeeles. Tartu Ülikool. Tartu. [Magistritöö].

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. Tartu Ülikool. [WWW] http://www.cs.ut.ee/~heli_u/magistritoo.html [Magistritöö].

Uibo, H. 2000a. Kahetasemeline morfoloogiamudel eesti keele arvutimorfoloogia alusena. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 37–72.

Uibo H. 2000b. On Using the Two-level Model as the Basis of Morphological Analysis and Synthesis of Estonian. – Proceedings from the 12th „Nordiske datalingvistikkdager" NODALIDA'99. Ed. by Nordgård. Department of Linguistics, NTNU. Trondheim, 228–242.

Uibo, H. 2002. Experimental Two-Level Morphology of Estonian. – Proceedings of the third International Conference on Language Resources and Evaluation. LREC 2002. Vol. 3. Las Palmas de Gran Canaria, Spain, 1012–1015.

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

Vider, K., Kahusk, N., Orav, H., Õim, H., Paldre, L. 2000. Eesti keele tesaurus. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 127–152.

Vider, K., Kaljurand, K. 2001. Automatic WSD: Does it make sense of Estonian? – Proceedings of SENSEVAL–2: Second International Workshop on Evaluating Word Sense Diasambiguating Systems. Toulouse, 159–162.

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

Viks, Ü. 2000a. Eesti keele avatud morfoloogiamudel. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. Hennoste, T. Tartu: Tartu Ülikooli Kirjastus, 9–36.

Viks, Ü. 2000b. Kuidas tekib sõnastikukirjesse grammatika. – Keel ja Kirjandus 7, 486–495.

Viks, Ü. 2000c. Tools for the Generation of Morphological Entries in Dictionaries. – Proceedings of the 2nd International Conference on Language Resources and Evaluation LREC2000. Athens.

Vossen, P. 1998a. EuroWordNet: Building a Multilingual Database with Wordnets for European Languages. – The ELRA Newsletter. Vol. 3. No 1. Ed. by Choukri, K., Fry, D., Nilsson, M.

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

Wilks, A.Y., Slator, B. M., Guthrie, L. M. 1996. Electric Words. Dictionaries, Computers, and Meanings. MIT Press.

Õim, H., Saluveer, M. 2002. Freimid keelekirjelduses. – Akadeemia 12, 2663–2682.

Trükitud
AS Pakett trükikojas