Eesti keele korpused ja arvutileksikonid — mis on olemas ja mida veel vaja on

23 aastat tagasi - 25.09.2002 Autor: AM

Autorid: **Kadri Muischnek**, **Heili Orav**

Sissejuhatus

Inimkeeled on mitmes mõttes sarnased, aga ka mitmes mõttes erinevad. Seega saab nende töötlemisel kasutada mingis osas keelest sõltumatuid algoritme ja programme, mingis osas aga on programm konkreetsest keelest sõltuv. Mõistlik viis üldist ja erilist ühendada on mõistagi selline, et konkreetse keelega seotud andmed esitatakse andmetena, keelest sõltumatud asjad aga programmina. Sel juhul saame näiteks Eesti keelt töödelda, asendades inglise keelele mõeldud programmis andmefaili (nt spelleri sõnastiku) Eesti omaga. Hea küll, spelleri puhul see niisama lihtsalt ei toimi, sest sõnamuutmise süsteem on liiga erinev. Mitmes muus, üldisemalt defineeritud valdkonnas (nt kõneanalüüs), on selline lähenemine siiski õigustatud: andmete hulka võivad kuuluda ka keelespetsiifilised reeglid, mida keelest sõltumatu tarkvara kasutab, keelematerjal ise, mida uurijatele „suupärasemaks” töötleb keelest sõltumatu tarkvara.

Kogu seda konkreetses keeles materjali, mida saab kasutada keelespetsiifiliste tarkvaratoodete valmistamiseks, eriti veel sellist, mis kõlbab mitmel erineval eesmärgil, nimetataksegi keeleressurssideks. Loomuliku kõne ja keele uurimisega tegelejad ongi jõudnud arusaamisele, et töökindlate ja tõhusate keeletoodete areng või keeletehnoloogiliste toodete areng sõltub otsustavalt sellest, kui kättesaadavad on suured adekvaatsed keeleressursid, nimelt elektroonilised sõnastikud, terminoloogiabaasid, teksti– ja kõnekorpused ning formaalsed grammatikad. Siin tahame rääkida Eesti keele tekstikorpustest ja arvutileksikonidest, nende hetkeseisust ja perspektiividest.

Korpused

Korpus on elektrooniline keele (teksti või kõne) kogum, mille alusel saab:

- analüüsida keelt, et tema omadusi kindlaks teha;
- treenida mingit arvutiprogrammi, et kohandada teda tööks tekstidega teatud piiritletud olukorras;

- empiiriliselt kontrollida keele kohta käivat teooriat;
- testida keeletehnoloogilist võtet või rakendust, et selgitada, kuidas ta töötab praktikas.

Ühte ja sama tekstikorpust saab tavaliselt kasutada mitmel erineval eesmärgil, nt sõnakasutuse, morfoloogia või grammatika uurimiseks, aga ka sõnastiku tegemiseks ja programmide testimiseks.

Korpused erinevad üksteisest nii mahu kui märgenduse detailsuse poolest. Märgendamine on vajalik, et muuta korpus paremini kasutatavaks. Nt tõlketekstide korpus, milles on iga lause juures viit originaalile, võimaldab koostada kakskeelseid sõnastikke ning luua uusi võimalusi pakkuvaid keeleõppe programme.

Korpuste tekstide valiku põhimõtete ja korpuste liigitusega tegeleb korpuslingvistika. Korpuslingvistilises teoorias on oluline mõiste korpuse representatiivsus ja balansseeritus. Selle all mõeldakse seda, kuidas on korpuses esitatud erinevad allkeeled (tekstiklassid). Näiteks tänapäeva Eesti keele (kirjutatud keele) korpuses on kümme tekstiklassi ja sõnade hulk igas tekstiklassis on proportsioonis selle tekstiklassi osatähtsusega kogu esindataval perioodil trükitud eestikeelses kirjasõnas. Olulisem on representatiivsus lingvistikale. Aga ka korpuste kasutamisel keeletehnoloogias tuleks jälgida, millistest tekstiliikidest kasutatav korpus koosneb, sest näiteks ilukirjanduse tekste kasutades treenitud programm ei suuda seadusetekste sama täpsusega analüüsida.

Eesti keele korpusi ja tekstikogusid koostatakse Eestis Tartu Ülikoolis ja Eesti Keele Instituudis. Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris on loomisel eestikeelse kõne andmebaas, mida saab kasutada näiteks tekst–kõne ja kõne–tekst süsteemide väljatöötamisel.

Tartu Ülikooli korpustele pääseb ligi veebilehekülgedelt www.murre.ut.ee ja www.cl.ut.ee Neist esimeses on sellised erikorpused nagu vana kirjakeele korpus ja Eesti murrete korpus ning suulise kõne korpus.

Arvutuslingvistika uurimisrühma koduleheküljel www.cl.ut.ee on kirjakeele (tähenduses kirjalik üldkeel) korpused. Nende keskmeks on 80ndate aastate korpus, mis sisaldab miljon sõna aastatest 1983–1987 ja koosneb kümnest tekstiklassist, mille osakaal korpuses esindab nende tekstide osakaalu kirjutatud keeles antud perioodil. Sellega liituvad nn läbilõikekorpused aastatest 1890–1990 — igast kümnendist on korpuses umbes 200 000 kuni 600 000 sõna ajakirjandus– ja ilukirjandustekste kui kirjutatud keele keskseid tekstiklasse. Selline korpuste süsteem annab hea ülevaate keele muutumisest 20. sajandi jooksul. Kuid

tänapäeva keelele rakendatavate keeletehnoloogiliste rakenduste väljatöötamise jaoks on enamus sellest materjalist muidugi lootusetult vananenud.

LINGID:

- www.phon.ioc.ee — Tallinna Tehnikaülikooli Küberneetika Instituudi foneetika ja kõnetehnoloogia labor
- www.hcu.ox.ac.uk/BNC — British National Corpus
- www.ids-mannheim.de/kt/corpora.shtml — Institut für Deutsche Sprache Mannheimis
- www.cogsci.princeton.edu/~wn — WordNet

Rääkides korpustest ei saa kuidagi mööda minna nende kasutajaliidestest, mis võimaldavad korpust kasutada. Tartu Ülikooli arvutuslingvistika uurimisrühma koduleheküljel asuvate korpuste kasutajaliides on kahetasemeline. Esimene võimaldab esitada päringuid sõna(vormi), sõna algvormi või morfoloogilise kategooria kohta. Lisaks sellele saab aga kasutada ka Unixi käsurida, käskude hulk on küll piiratud käskudega: cut, grep, head, join, paste, rev, sed, sort, tail, tr, uniq, wc, kuid nende kombinatsioonid annavad korpuse kasutajale võimaluse teha täpselt sellist päringut, nagu ta tahab.

Teine (praegu suurem kui Tartu Ülikooli oma) Eesti keele korpus on koostatud Eesti Keele Instituudis: www.eki.ee/corpus. Selle korpuse maht on 10,4 miljonit sõnavormi, umbes 80% ulatuses ajalehetekstid. Ka see korpus on varustatud kasutajaliidesega, mis võimaldab teda üle Interneti kasutada. Korpus on koostatud juhuslikult kogutud materjalist ja pole seega representatiivne. Ka ei ole korpus märgendatud, sobides eelkõige leksikaalse materjali otsinguks.

Siiani oli juttu sellest, mis meil olemas on. Aga mida veel vaja oleks? Lingvistika jaoks, keeleuurimiseks läheb vaja suuremaid, kuid siiski representatiivseid korpusi. Tõeliselt suuri representatiivseid korpusi on maailmas koostatud suhteliselt vähe, üks tänapäevasemaid näiteid on briti inglise keele British National Corpus.

Keeletehnoloogia vajadused võivad olla veidi teistsugused kui lingvistikal. Samas tuleb rõhutada, et mida keerulisem on mingi keeletehnoloogiline rakendus, seda lihtsam on teha teda pigem mingi kitsa allkeele kui kogu üldkeele jaoks. Piltlikult ja veidi liialdades: on võimalik luua masintõlkesüsteem, mis tõlgib edukalt Hewlett–Packardi printeri manuaale, kuid mingi teise firma manuaalidega võib ta juba hätta jääda. Sellise süsteemi loomist alustatakse muidugi kõigi Hewlett–Packardi printeri manuaalide korpuseks koondamisega.

Seda probleemi on mujal maailmas lahendatud tavaliselt sellel teel, et korpuste mahtu on suurendatud nii palju, et iga kasutaja saab sellest teha valikuid vastavalt oma vajadustele. Piltlikult: sealt leiab nii printeri manuaalid kui Tammsaare teosed. Tõeliselt suure korpuse näitena võiks tuua seekord mitte inglise, vaid hoopis saksa keele korpuse, mida on kogunud Mannheimis Institut für Deutsche Sprache, ja mille suurus on miljard 83 miljonit sõna. Selle korpuse struktuur näitab ilmekalt korpuslingvistika suundumusi: kogutakse seda, mida saab koguda suhteliselt lihtsalt — sellest ajakirjanduse väga suur osakaal. Ilukirjandustekstide korpusesse viimisel on nimelt alati suured probleemid autoriõigusega, aga saada kirjastuselt lepingut, mis lubab kasutada juba ilmunud ajalehti, on tunduvalt lihtsam.

Selliseid tõeliselt suuri korpusi ei ole ainult suurte keelte jaoks nagu inglise ja saksa, vaid selleni on jõutud ka tunduvalt väiksema kõnelejate arvuga keelte uurimisel. Nii on sloveenidel 100 miljonist sõnast koosnev korpus (www.fida.net/eng/index.html), leedukate korpuses (donelaitis.vdu.lt/) on praegu 60 000 sõna ja nad kavatsevad seda veelgi suurendada. Ka eesti keele uurimine ja eriti Eesti keeletehnoloogia ei saa kuidagi läbi tõeliselt suure korpuseta, mille koostamisega tegeleb praegu Tartu Ülikoolis projekt „ Eesti keele segakorpus”, mida on seni rahastatud riiklikust sihtprogrammist „Eesti keel ja rahvuskultuur”.

Projekti eesmärgiks on piisavalt suure (ideaalis 100 miljonit sõna) tekstikorpuse kogumine; olemasolevate ja uute tekstide üleviimine ühtsele märgendussüsteemile, korpuse ülespanek internetti ja hooldus. Korpusesse otsustasime võtta ainult terviktekstid, mitte tekstikatked, millest suures osas koosneb Tänapäeva Eesti Kirjakeele Korpus.

Korpuslingvistikas räägitakse palju korpuse representatiivsusest, mis tähendab seda, et korpuses peaksid olema esindatud kõik (või valitud) tekstiklassid, mis antud kultuuris antud ajavahemikul olemas on ja korpuse balansseeritusest, mis tähendab seda, et nende tekstiklasside esindatus korpuses peab vastama nende esindatusele antud kultuuris. tegelikult kaotavad representatiivsus ja balansseeritus oma tähtsust sedamööda, kui korpused järjest mahukamaks muutuvad. Selle korpuse korjamisel me (vähemalt praegu) ei tegele tekstide tekstiklassideks jaotamisega ega jälgi proportsioone tekstiklasside vahel. Korpuse loodame ajada nii suureks, et kui keeleuurijal on vaja representatiivsust ja balansseeritust, siis saab ta teha olemasolevatest tekstidest omad valikud. Esialgu kogume ainult seda, mida saame elektroonilisel kujul. Kui selgub, et mingi oluline tekstiklass (nt ilukirjandus) on selgelt alaesindatud, siis tuleb mõelda tekstide skannimisele.

Elektroonilised sõnastikud

Iga rakendus, mis kasutab sõnu, vajab ka arvutisõnastikke ning tihti on just sõnaraamat süsteemi kerkseks osaks. Elektroonilised sõnastikud erinevad traditsioonilistest, inimese jaoks mõeldud (paber)sõnastikest nii oma struktuuri kui sisu poolest. Kõik arvutirakendused ei vaja 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. Oluline on keeletehnoloogiat kasutavate sõnaraamatute loomine (algvormide ja mitmesõnaliste fraaside automaatne leidmine tekstist). Sellised sõnaraamatud oleksid ka samm tõlkijate abivahendite ning masintõlke suunas. Elektrooniliste sõnaraamatute uued põlvkonnad juba sisaldavad keeletehnoloogia elemente (algvormide leidmist, fraaside automaatset leidmist tekstist jm).

Areng arvutileksikograafia sees on kulgenud arvutisse sisestatud sõnastikutekstidelt leksikaalsete teadmusbaasideni:

Arvutiversioon

Eestis on paljud pabersõnastikud antud välja ka elektrooniliselt. Hulk elektroonilisi sõnastikke on kasutatavad ka Internetis. 1998. aastal loodi KeeleWeb (ee.www.ee), mis on paljusid valminud sõnastikke ja keeletehnoloogia mooduleid ühendav keskkond. Võrgus on väljas järgmised Eesti keele sõnastikud:

- õigekeelsussõnaraamat (ÕS76 täiendatud variant),
- antonüümisõnastik (Õim, A. 1995),
- sünonüümisõnastik (Õim, A. 1991),
- fraseoloogiasõnaraamat (Õim, A. 1993),
- tesaurus (Filosoft),
- slängisõnastik (Loog 1991),
- murdesõnastik („Väike murdesõnastik”, 1982–1988),
- inglise– Eesti –inglise (IBS),
- inglise– Eesti (Festart),
- inglise–vene (Festart),
- Eesti –vene– Eesti (ASE)

Momendil on suureks probleemiks põhjaliku tänapäevase elektroonilise inglise– Eesti ja Eesti –inglise sõnastiku puudumine.
Inglise– Eesti masintõlkesõnastiku loomisega tegeleb EKI. Selle projekti 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 70 000) (www.eki.ee/keeletehnoloogia/projektid/inglise-eesti).

Arvutisõnastik ei ole siiski vaid arvutisse viidud sõnaraamatu tekst. Sõnastikuartikli erinevad funktsionaalsed osad (märksõna ise, grammatiline info, seletus, näited) peavad olema formaalselt identifitseeritavad, nt varustatud spetsiifiliste märgenditega. Just tänu sellisele liigendusele on sõnastikus esitatud materjal ka „arvuti poolt loetav" ning mitte ainult inimese poolt kasutatav raamatu asendajana. Arvuti abil võidakse otsida või analüüsida sõnastikuartikli erinevaid osi eraldi, nt seletusi, näiteid, grammatilist infot. Pabersõnastike arvutiversioone kasutatakse leksikaalsete andmebaaside ja teadmusbaaside tegemiseks, aga ka igasuguse muu lingvistilise info otsinguks.

Leksikaalne andmebaas

Leksikaalse andmebaasi 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 liik mõistelist sõnaraamatut, 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.

1996. aastaks oli selge, et lisaks Eesti keele morfoloogia ja süntaksi arvutile arusaadavaks tegemisele on edaspidi vaja ka sõnasemantikal põhinevat leksikaalset andmebaasi. Mujal maailmas teostatud semantiliste arvutileksikonide

seas ringi vaadates tundus sobivaim olevat WordNeti idee.

WordNet (WN), mille loomist alustati 1980–ndate aastate keskel, 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 andmebaas, vaid tähenduste järgi.

WN–i elementaarosake on sünonüümirida — sünohulk (synonym set, synset), mille moodustavad ühte mõistet väljendavad sünonüümsed sõnad ja sõnaühendid. Termin sünohulk on loodud sellepärast, et erinevalt sünonüümisõnastiku sünonüümireast võib sünohulk olla ka üheliikmeline. Kui sünonüümisõnastiku eesmärgiks on kõigi võimalike keeles leiduvate sünonüümide esitamine, siis WN–i eesmärgiks on mõistete esitamine, ka siis, kui selle väljendamiseks keeles leidub ainult üks leksikaalne üksus.

Näide: sünohulk = inimene, inimolend, indiviid, isik, persoon, hingeline, hing — olend, keda iseloomustab kõrgelt arenenud aju, abstraktse mõtlemise võime ja artikuleeritud kõne, homo sapiens. Loomad võivad niimoodi käituda, mitte inimesed.

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

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

Eestis alustati EuroWordNet–i (EWN) projekti raames Eesti üldkeele tesauruse ehk Eesti wordneti (EstWN) koostamist 1997.a TÜ arvutuslingvistika uurimisrühmas. EWN (www.illc.uva.nl/EuroWordNet) oli Euroopa Komisjoni projekt aastatel 1996–1999, mille eesmärgiks oli luua WN–i eeskujul mitmekeelne leksikaal–semantiline andmebaas, milles erinevate keelte (inglise, hollandi, itaalia, hispaania, prantsuse, saksa, tšehhi, eesti) wordnetid on ühendatud.

EWN peamine erinevus WN–st ongi tema mitmekeelsus. Kõik projektis osalejad lõid WN–i põhimõttelisele ülesehitusele toetudes omakeelse wordneti, kuid keeltevahelise indeksi (interlingual index, ILI) kaudu on võimalik leida sama

mõistet väljendavad sünonüümihulgad teistes keeltes.

Sünohulki on EstWN–s hetkel kümne tuhande ringis — põhiliselt substantiivi– (66%) ja verbimõisted (27%), kuid vähesel hulgal ka adjektiive (2,6%) ja pärisnimesid (4,4%). Tesauruse kasv pole olnud nii kiire, kui me algselt lootsime. Praegune siht on saavutada 15–tuhande mõistega tesaurus, mida plaanime panna ka Internetti.

Leksikaalselt põhineb loodav tesaurus olemasolevatel traditsioonilistel sõnaraamatutel (peamiselt Eesti Kirjakeele Seletussõnaraamatul) ja tekstikorpusel (mis annab teavet sõnakasutusest), seega võib semantilist informatsiooni, mida andmebaas sisaldab, pidada keelelisel teadmisel põhinevaks.

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

Alates 1996. aastast on Eesti üldkeele tesauruse loomist toetanud Eesti Teadusfond ja Eesti Informaatikakeskus sihtprogrammis „ Eesti keeletehnoloogia”. Samuti riikliku sihtprogrammi „Eesti keel ja rahvuskultuur” keeletehnoloogia allprojekt.

Sarnaseid leksikaalsemantilisi ressursse saab kasutada automaatsete tõlkesõnastike ja intelligentsete info–otsisüsteemide, mis on võimelised otsima mõisteid või tähendusi mitmetes erinevates keeltes, loomiseks.

Leksikaalsed teadmisbaasid

Viimasel ajal on leksikaalsete andmebaaside kõrval üha enam hakatud rääkima ka leksikaalsetest teadmisbaasidest. Üks peamisi erinevusi leksikaalsete teadmisbaaside ja leksikaalsete andmebaaside vahel on esimeste võime esile tuua üldistusi ja tuletada järeldusi. 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. Seega on see maailmateadmistel põhinev andmebaas, kuna leksikaalses andmebaasis on ainult keelelised teadmised.

Missuguse teoreetilise mudeli raames ja missuguste tehniliste vahenditega leksikonile sellised omadused tagada, on aga täielikult veel diskussioonide objekt

kogu maailmas.

Kokkuvõte

Elektrooniliste sõnastike tegemisega on rohkem või vähem tegevuses TÜ, EKI, Festart, Filosoft jm uurimis– ja kommertsasutused.

Arvutileksikoni loomine eeldab — nii nagu iga teinegi arvutuslingvistiline või keeletehnoloogiline rakendusülesanne — mitme eriala inimeste koostööd. On vaja inimesi, kes oleksid piisavalt kompetentsed keeleteoorias, leksikoloogias ja leksikograafias, keelekirjelduse formalismides, korpuste kasutamises, arvutuslingvistikas ja arvutiteaduses. Võib–olla just sellistest spetsialistidest on meil kõige suurem puudus.

Lõppsõna

Keeleressursid on keeletehnoloogia jaoks elulise tähtsusega. Nende koostamisse tasub investeerida aega ja raha. Samuti oleks hea, kui firmadele ja ka uurijatele oleksid need lingvistilised ressursid kättesaadavad — sõnastikud arvutis, formaalsed grammatikad, tekstide kogumid ehk korpused. Nad saaksid nende hulgast valida mingi alamhulga, mida konkreetsete operatsioonisüsteemide, arvutite ja standarditega seotud programmide aluseks võtta.

Artikli aluseks on Tartus 27. juunil 35. Johannes Voldemar Veski päeval peetud ettekanne.

KIRJANDUS:

1. Hennoste, T. Tartu University Corpus of Written Estonian: A Survey of the Structure of Texts and Principles of Selection. Kogumikus: Estonian in the Changing World. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised. Tartu 1996, lk 7–32
2. Hennoste, T., Kaalep, H–J., Muischnek, K., Paldre, L., Vaino, T. The Tartu University Corpus of Estonian Literary Language. Congressus Nonus Fenno–Ugristarum Pars V, Tartu 2001, pp 337–344.
3. Hennoste, T., Muischnek, K. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toimetaja Tiit Hennoste. Tartu 2000, lk 183–218.
4. Hennoste, Tiit, Liina Lindström, Andriela Rääbis, Piret Toomet, Riina Vellerind 2000. Eesti suulise kõne korpus ja mõnede allkeelte võrdluse katse. — rmt T. Hennoste (toim.) Arvutuslingvistikalt inimesele, Tartu Ülikooli üldkeelteaduse

õppetooli toimetised 1. Tartu, lk 245–283.
5. Hennoste, T., Koit, M., Kullasaar, M., Rääbis, A., Vutt, E., Eesti dialoogikorpuse loomise probleemidest. — rmt T. Hennoste (toim.) Arvutuslingvistikalt inimesele, Tartu Ülikooli üldkeelteaduse õppetooli toimetised 1. Tartu, lk 143–160
6. Koit, M., Roosmaa, T. Overview of the Uses of the Corpus of Literary Language: Current Possibilities. Kogumikus: Estonian in the Changing World. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised. Tartu 1996, lk 33–42
7. Langemets, M. Sõnaraamatu arvutilingvistiline analüüs. Magistritöö. EKI, Tallinn, 2000, lk. 20–21
8. Vider, Kadri ja Orav, Heili 1998. Sõna tasandilt mõiste ruumi. „Keel ja Kirjandus", nr. 1, 1998, lk. 57–64
9. Vider, Kadri; Kahusk, Neeme; Orav, Heili; Õim, Haldur; Paldre, Leho 2000. Eesti keele tesaurus. — Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T. Hennoste. Tartu. lk. 127–152.

Lingid samal teemal:

www.murre.ut.ee
www.cl.ut.ee
www.eki.ee/corpus
www.fida.net/eng/index.html
donelaitis.vdu.lt/
ee.www.ee
www.eki.ee/keeletehnoloogia/projektid/inglise-eesti
www.illc.uva.nl/EuroWordNet
www.phon.ioc.ee
www.hcu.ox.ac.uk/BNC
www.ids-mannheim.de/kt/corpora.shtml
www.cogsci.princeton.edu/~wn

- Lahendused