Keel ja Kirjandus 1/1998

EESTI TEADUSTE AKADEEMIA JA EESTI KIRJANIKE LIIDU AJAKIRI XLI AASTAKÄIK

ARVUTUSLINGVISTIKA MUJAL JA MEIL

MARE KOIT, HALDUR ÕIM

Mis on arvutuslingvistika?

Arvutuslingvistika on keeleteaduse ja arvutiteaduse ehk informaatika vahepeal asuv hübriidala, mis tegeleb loomuliku keele automaattöötluseks vajalike keele kirjeldus-, analüüsi- ja sünteesimeetodite väljatöötamise ja arvuti abil realiseerimisega.[1] Arvutuslingvistikal on seoseid ka matemaatikaga, psühholoogiaga ning inimese intellektuaalset tegevust uuriva ja modelleeriva intellektitehnikaga (tehisintellektiga). Arvutuslingvistika kombineerib humanitaar-, loodus- ja tehnikateaduste meetodeid. Arvutuslingvistikas saab eristada teoreetilist ja rakenduslikku poolt.[2] Teoreetilise komponendi sisuks on inimese keelepädevuse kohta käivate teooriate püstitamine ja kontrollimine. Rakenduslik komponent keskendub inimese keelekasutuse modelleerimise praktilisele väljundile; vastavad meetodid, tehnikad, tööriistad ja rakendused moodustavad nn keeletehnoloogia.[3]

Loomuliku keele töötlemisel on üritatud arvuteid kasutada juba alates 1940-ndate aastate lõpust. Algul keskenduti põhiliselt masintõlkele (vene keelest inglise keelde). Esimestes masintõlkesüsteemides alahinnati teoreetilisi raskusi, edaspidi on pööratud teoreetilisele lingvistikale rohkem tähelepanu. Areng ongi kulgenud tsüklitena, kus põhirõhk on olnud kord teoreetilisematel, kord praktilisematel probleemidel.[4] Viimastel aastatel on kõrvuti teooriaga taas intensiivselt arenenud rakenduslik suund, keeletehnoloogia. Interneti ja

[1] Vt ka: H. Õim, Inimene, keel, arvuti. Tallinn, 1983; M. Koit, Arvutuslingvistika arenguloost ja stuudiumist. — Keel ja Kirjandus 1996, nr 3, lk 171—178.

[2] M. Koit, Arvutuslingvistika arenguloost ja stuudiumist, lk 171.

[3] Vahel määratletakse arvutuslingvistikat ainult kui teoreetilist teadust, lugedes rakendusi eraldiseisvateks uurimisvaldkondadeks (rakenduslingvistika, korpuslingvistika, grammatika- ja keeletehnoloogia). Siin mõistame arvutuslingvistikat siiski laiemalt.

[4] Vt ka: M. Koit, Arvutuslingvistika arenguloost ja stuudiumist, lk 171—175.

infosüsteemi WWW — veebi — kiire areng ning mitmekeelsus seavad keeletehnoloogia ette täiendavaid ülesandeid. Kasutajasõbraliku tarkvara (s.t arvutiprogrammide) järele on üha kasvav vajadus. Inimese tööd lihtsustavad juba praegu õigekirja- ja grammatikakontrollijad tekstitoimetite koosseisus, intelligentsed elektronkirjade filtreerijad ja edasitoimetajad, tekstide klassifitseerimise süsteemid, automaatse refereerimise süsteemid jms. Viimasel ajal on muutunud kättesaadavaks kõne analüüsi ja sünteesi vahendid alates pimedatele mõeldud süsteemidest, mis teisendavad teksti kõneks, kuni kontorites kasutatavate dikteerimissüsteemideni, mis teisendavad kõne tekstiks.

Arvutuslingvistika hetkeseis maailmas

Vahendid, valdkonnad ja probleemid.[5] 1. Keeleressursid. Keeleehk lingvistiliste ressursside all mõistetakse suuri masinalloetavaid keeleandmete hulki ja kirjeldusi, mida saab kasutada teksti- ja kõnetöötlussüsteemide ülesehitamisel. Keeleressurssideks on näiteks teksti- ja kõnekorpused, leksikaalsed andmebaasid, grammatikad, arvutileksikonid, aga ka selliste ressurssside kogumiseks ja kasutamiseks vajalik tarkvara. 1995. aastal asutati Luxembourgis Euroopa Lingvistiliste Ressursside Assotsiatsioon (European Linguistic Resources Association, ELRA; *http://www.de.relator.research.ec.org/elra/*), mille ülesandeks on Euroopa keelte lingvistiliste ressursside kogumine, nende edasiarendamine, haldamine ja levitamine. 1997. aasta lõpuks on olemas 64 kõneandmebaasi mitme keele jaoks, 15 ühe- või mitmekeelset tekstikorpust, 40 üheja 60 mitmekeelset leksikoni, lingvistilise tarkvara ja grammatikaarenduse platvormid, üle 360 terminoloogilise andmebaasi mitme ainevaldkonna ja paljude keelte jaoks. USA-s on analoogiliseks organisatsiooniks Lingvistiline Andmekonsortsium (Linguistic Data Consortium, LDC; *http://www.ldc.upenn.edu/*).

2. Matemaatilised meetodid. Morfoloogias, fonoloogias ja süntaksis kasutatakse formaalsete keelte teooriast lähtuvaid matemaatilisi meetodeid, mille aluseks on N. Chomsky formaalsete grammatikate ja keelte hierarhia ning viimasel kümnendil eeskätt unifikatsioonigrammatikaid. Nendes on rakendatud spetsiifilisi töötlusmeetodeid, mis on tihedalt seotud kitsendustega, loogilise programmeerimisega. Semantikas vajatakse esituskeeli, millega üheselt väljendada tähendust. Selleks rakendatakse loogikat, eeskätt kõrgemat järku predikaatarvutust. Mõningad esituskeeled, näiteks freimid ja skriptid, on üle võetud intellektitehnikast. Tuletustehnikana kasutatakse põhiliselt loogilist deduktsiooni. Kuigi lingvistilise teadmuse esitus on tänu uutele esitustehnikatele muutunud adekvaatsemaks, on olemasolevad süsteemid siiski veel vähese tõrkekindluse ja efektiivsusega. Seetõttu on pöördutud statistiliste meetodite poole. Kõnetuvastuses on ammu kasutatud spetsiaalset tüüpi tõenäosuslikke automaate, Markovi varjatud mudeleid. Praegu kasutatakse statistilisi meetodeid peaaegu igas keeletöötluse valdkonnas (statistiline sõnaliikide määramine, tõenäosuslik süntaksianalüüs, morfoloogilise ja süntaktilise mitmesuse lahendamine, leksikaalsete teadmiste hankimine, statistiline masintõlge). Rakendatakse optimeerimistehnikaid ja konnektsionistlikke meetodeid, näiteks kõnetuvastuses närvivõrkude meetodit.

Kolmas on "madala taseme", otse täitmisele orienteeritud meetodite klass, nimetatagu näiteks lõplikke automaate või laiendatud üleminekuvõrke kasutavat analüüsi (*Augmented Transition Network*, ATN).

3. Kõne analüüs ja süntees. Kõne abil arvuti poole pöördumine on probleem, millega insenerid ja uurijad on tegelnud juba peaaegu viis aastaküm-

[5] Vt Survey of the State of the Art in Human Language Technology, 1996. *http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html.*

met. Praeguseks on kõnetuvastustehnoloogias hakatud laboratoorsetelt demoversioonidelt üle minema kommertsrakendustele. Kõnetuvastussüsteemid jagunevad isoleeritud sõna ja seotud kõne tuvastamise süsteemideks. Teine tegur, mis eristab kõnetuvastussüsteeme, on sõnastiku maht. Siin on esindatud äärmused: ainukõneleja 30 sõnast kuni erinevate kõnelejate 10 000 sõnani. Kõnesisendi puhul on süsteemil tegu ebakindlusega, mida kirjutatud sõnade puhul ei esine. Lisaks on kõne ka struktuurilt teistsugune. Rääkimisel produtseeritavad häälikud muudetakse digitaalseks. Seejärel töödeldakse seda signaali, et elimineerida mitmesuguseid tunnuseid, nagu hääliku intensiivsus erinevatel sagedustel ja intensiivsuse muutumine aja jooksul. Need tunnused on kõnetuvastussüsteemi sisendiks. Üldiselt kasutab selline süsteem Markovi varjatud mudeli tehnikat määramaks kõige tõenäolisemat sõnade järjendit, millest moodustub väljund. See on omakorda loomuliku keele teksti mõistmise süsteemi sisendiks.

Kõnesünteesis püüti algul enamasti modelleerida inimese kõneloomemehhanisme; praegu kasutatakse lisaks ka muid lähenemisviise. Uus tehnoloogia — tekstist kõne sünteesimine — aitab parandada sünteesitava kõne kvaliteeti. Vokaalsete karakteristikute kohandamine võimaldab väljundteksti varieerida.

4. Teksti analüüs ja süntees. Teksti analüüs koosneb traditsiooniliselt kolmest etapist: morfoloogilisest, süntaktilisest ja semantilisest.

Morfoloogilise analüüsi automatiseerimisel on viimase kümmekonna aastaga saavutatud olulist äriedu. Hästi tuntud on Kimmo Koskenniemi kahetasemeline mudel, mida on rakendatud mitme keele puhul (inglise, prantsuse, soome, kreeka jpt). Morfoloogilise analüsaatori väljundi ühestamisel (s.t võimalike analüüsivariantide hulgast ühe, õige valikul) kasutatakse kaht põhimeetodit: reeglitel põhinevat ja tõenäosuslikku. Esimesel juhul asetatakse põhirõhk ühestamise täpsusele, teisel aga täielikkusele (kas või vigade hinnaga). Näiteks Fred Karlssoni kitsenduste grammatikal[6] põhinev inglise keele ühestaja kasutab 1100 reeglit ja selle täpsus on 99,7%; ühestamata jääb 2—6% sõnadest. Tõenäosuslikud ühestajad annavad inglise keele puhul täpsuseks 96%. Perspektiivikaks peetakse kahe meetodi kombineerimist.

Süntaksianalüüsis kasutatakse eeskätt mitmesuguseid unifikatsioonigrammatikaid, kus on mitusada kuni mitu tuhat reeglit. Levinumad on peajuhitav fraasistruktuurigrammatika (*Head-Driven Phrase Structure Grammar*, HPSG), funktsionaalne unifikatsioonigrammatika (FUG), leksikaalfunktsionaalne grammatika (LFG) ja puulaiendamisgrammatika (*Tree Adjoining Grammar*, TAG). Mitmes praegu USA-s, Kanadas, Jaapanis, Lääne-Euroopas ja Austraalias väljatöötatavas loomuliku keele süsteemis kasutatakse HPSG formalismi[7]. Suurim neist on *Verbmobil*, mille Saksamaa valitsus algatas 1993. aastal ja kus osaleb üle 30 organisatsiooni. Teisiti läheneb süntaksianalüüsile eespool nimetatud kitsenduste grammatika, hõlmates nii süntaksianalüüsi kitsamas mõttes kui ka morfoloogilist ühestamist selle eeletapina. Kitsenduste grammatika koosneb paljudest üksteisest enamasti sõltumatutest reeglitest, millest igaüks esitab mõne keelereeglilaadse fakti. Need grammatilised reeglid — kitsendused (ingl *constraints*) — ei määra lause grammatilist korrektsust, nagu teevad seda paljud teiste grammatiliste formalismide reeglid, vaid püüavad leida morfoloogiliste tunnuste ja konteksti info põhjal, missugust funktsiooni konkreetne sõna lauses täidab.

Teksti genereerimisel on põhiprobleemideks teksti ja lausete planeerimine

[6] Constraint Grammar. A Language-Independent Formalism for Parsing Unrestricted Text. Berlin—New York, 1995.

[7] Vt *http://hpsg.stanford.edu/hpsg/hpsg.html*.

ning plaani grammatiliselt korrektseks tekstiks teisendamine. Loomuliku keele süsteemides kasutatakse sageli lihtsaimat lähenemisviisi: süsteem väljastab valmis lauseid (veateateid, hoiatusi). Järgmine tase on šabloonide kasutamine, kui teadet tuleb produtseerida korduvalt, kuid väikeste muudatustega. Täiuslikumad süsteemid kasutavad tunnustel põhinevat lähenemist, kus väljund ehitatakse üles lihtsatest tunnustest struktuuri moodustamise teel. Nii genereeritakse praegu üksiklauseid, mitte veel seotud tekste. Mitmes kohas, näiteks Saarbrückenis Saksamaa Tehisintellekti Uurimiskeskuses ja Ameerikas Columbia ülikoolis on teoksil tööd, kus rakendatakse multimeediat koos planeerimise ja keele genereerimisega. Siiani pole siiski suudetud välja töötada ainevaldkonna modelleerimise, diskursuse struktuuri ja lause planeerimise ega sõnavaliku üldisi meetodeid. Abi loodetakse keeleressurssidest, sh suurtest hästi struktureeritud leksikonidest ja grammatikatest. Ameerika Ühendriikides on loodud üldeesmärgiline teadmusbaas (semantiline sõnastik) WordNet[8], mida saab kasutada valdkonnaspetsiifiliste rakenduste konstrueerimisel. Euroopas on loomisel selle mitmekeelne analoog EuroWordNet.

Organisatsioonid, publikatsioonid ja üritused. Tähtsaimaks rahvusvaheliseks organisatsiooniks on 1968. aastal loodud Arvutuslingvistika Assotsiatsioon (Association for Computational Linguistics, ACL; *http://www.cs.columbia.edu/~acl*), mille enamik liikmeid on USA-st ja Euroopast. Assotsiatsioon annab välja ajakirja Computational Linguistics ja hooldab elektroonilist arhiivi "Computation and Language E-Print Archive" (*http://xxx.lanl.gov/cmp-lg/*). Organisatsioonil on ka Euroopa osakond (European Chapter, EACL).

1991. aastal loodi Amsterdamis Euroopa Loogika, Keele ja Informatsiooni Assotsiatsioon (FoLLI), mille eesmärgiks on arendada teadustööd ja haridust loogika, keeleteaduse, arvutiteaduse ja kognitiivteaduse alal ning nendega piirnevates distsipliinides (*http://www.wins.uva.nl/research/folli/*). Ühtlasi toetatakse elektroonilise ajalehe COLIBRI väljaandmist (*http://colibri.let.ruu.nl*), see ilmub iga nädal ja edastab loomuliku keele töötlemise ja loogika valdkonna uudiseid (konverentsid jm üritused, töö- ja koolituspakkumised, raamatute ja tarkvara esitlused).

1991. aastal asutati ka Euroopa Keelte ja Kõne Võrk (European Network in Language and Speech, ELSNET; *http://www.elsnet.org/*), kuhu kuulub praegu 80 akadeemilist ja 47 tööstusorganisatsiooni. Selle ülesandeks on toetada keeletehnoloogia arengut Euroopas, mitmekeelsete loomuliku keele süsteemide loomist.

Tähtsamad ajakirjad on eespool juba mainitud Computational Linguistics, Computer Speech & Language (orienteeritud kõnetöötlusele), Machine Translation (masintõlkealane), Journal of Natural Language Engineering, Mind and Language, Journal of Logic, Language and Information. Peatselt hakkab ilmuma (sagedusega kord kvartalis) elektrooniline ajakiri Journal of Language and Computation, mille väljaandmist toetab FoLLI.

Arvutuslingvistikaga tegelevatel ülikoolidel ja institutsioonidel on ka veebilehekülgi, nt Zürichi ülikooli *http://www.ifi.unizh.ch/CL/*, Koblenz-Landau ülikooli *http://www.uni-koblenz.de/~compling/*, Manchesteri ülikooli teaduse ja tehnoloogia instituudi *http://www.umist.ac.uk/* jpt. On spetsiaalseid elektroonilisi postiloendeid, nt *linguist@tamsun.tamu.edu, elsnet-list@let.ruu.nl, tei-l@uicvm.bitnet*, samuti uudisegruppe, nt *comp.ai.nat-lang*, kus on järgmised alajaotused: loomuliku keele mõistmine, genereerimine, masintõlge, dia-

[8] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, Introduction to Wordnet. An On-Line Lexical Database. — International Journal of Lexicography 1990, kd 3, nr 4, lk 235—244.

loogi ja diskursuse süsteemid, loomuliku keele liidesed, süntaksianalüüs, arvutuslingvistika, arvutitoetusega keeleõpe.

Traditsiooniliselt korraldatakse rahvusvahelisi üritusi, millest esinduslikem on 1965. aastast iga kahe aasta järel toimuv COLING. Igal aastal viiakse läbi Rahvusvahelise Arvutuslingvistika Assotsiatsiooni (ACL) aastakoosolek. 1998. aasta augustis toimuvad Kanadas Montreali ülikoolis ühisüritusena COLING '98 (arvult seitsmeteistkümnes) ja ACL '98 (kolmekümne kuues). Oluline üritus on veel keele, loogika ja informatsiooni alane Euroopa suvekool ESSLLI (European Summer School of Language, Logic and Information), mille läbiviimist toetab FoLLI; kümnes selline kool toimub 1998. aasta augustis Saarbrückenis.

Õpetamine ja tööturg. Arvutuslingvistikat hakati maailma ülikoolides iseseisva erialana õpetama 1980-ndatel aastatel;[9] eriala omandanud noored leiavad tööd ülikoolides, riiklikes uurimisasutustes või suurtes ettevõtetes. Näiteks Saksamaal töötavad arvutuslingvistid uurimisrühmades ülikoolide juures, kus seda eriala õpetatakse (Bielefeld, Koblenz, Saarbrücken, Stuttgart), või riiklikes uurimislaboratooriumides, nagu Matemaatika ja Andmetöötluse Ühing (Gesellschaft für Mathematik und Datenverarbeitung, GMD) või Tehisintellekti Uurimiskeskus, ning kompaniides, nagu Siemens või IBM. Lisaks neile on arendusgrupid, mis töötavad välja kommertstooteid. Üldiselt on nii teoreetilise kui ka praktilise kallakuga arvutuslingvistide järele pidev (kuigi mitte eriti suur) vajadus.

Arvutuslingvistika Eestis

Eestis tegeldakse arvutuslingvistikaga põhiliselt Tartu ülikoolis ja Tallinnas Eesti Keele Instituudis, keeletehnoloogiaga ka Küberneetika Instituudis.

Ajaloost. Eestis võib arvutuslingvistika alguseks lugeda 1950-ndate aastate lõppu, kui Tartu ülikooli matemaatika-loodusteaduskonnas alustati Ülo Kaasiku juhendamisel masintõlkealast tööd (eesmärgiks oli tõlkida matemaatilisi tekste vene keelest eesti keelde). Arvutile Ural-1 koostati vene keele morfoloogilise analüüsi programm, mis kasutas 2500 sõnast koosnevat sõnastikku. 1960-ndate aastate keskel jäi see töö aga soiku.

1965. aastal moodustati Tartu ülikooli eesti keele kateedri juures strukturaallingvistika töörühm (hilisema nimetusega generatiivse grammatika grupp ehk GGG), mida juhendas Huno Rätsep. Rühma kuulus seitsme tööaasta jooksul ligi 20 inimest. Algul oli tähelepanu keskmes generatiivse grammatika teooria, millest lähtudes rakendati eesti keele uurimiseks uusi analüüsi- ja kirjeldamismeetodeid. Anti välja sarja "Keel ja struktuur" (kokku 10 köidet). 1966. aastal hakkas rühm avaldama TÜ toimetistes pikemate uurimuste sarja "Keele modelleerimise probleeme", millest ilmus seitse köidet. Selles rühmas alustas oma teadustööd mitu praegu arvutuslingvistikas tegutsevat teadlast, ka mõlemad siinse kirjutise autorid osalesid rühma töös.

1960-ndate aastate lõpul moodustati TÜ-s juristidest, keeleteadlastest ja matemaatikutest uurimisrühm, kes prof Ilo Sildmäe (1922—1992) juhendamisel töötas välja tesaurusel põhineva süsteemi juriidilise info otsimiseks (kasutati arvutit Minsk-32). Aastatel 1978—1991 andis uurimisgrupp välja TÜ toimetiste sarjas tehisintellektialaseid kogumikke (kokku 12 numbrit). 1980. aastal rajati TÜ tehisintellekti labor, milles tegutses kaks uurimisrühma. I. Sildmäe juhendamisel töötati välja tema teaduslikele ideedele (teadmiste olemuse selgitamine, teadmiste esitamine) tuginevat eksperimentaalset süsteemi, tegeldes seejuures venekeelse teksti automaattöötlusega. Teist

[9] Vt M. K o i t, Arvutuslingvistika arenguloost ja stuudiumist, lk 176—178.

rühma juhendas Haldur Õim. Rühm tegeles eesti keele automaatse morfoloogilise, süntaktilise ja semantilise analüüsi ja sünteesiga, seotud teksti mõistmise ning arvuti ja inimese vahelise dialoogi modelleerimisega. Arvutil realiseeriti eksperimentaalne süsteem. Pärast seda kui 1993. aastal TÜ struktuuri muudeti, jätkub arvutuslingvistikaalane töö üldkeeleteaduse õppetooli juurde loodud arvutuslingvistika uurimisgrupis prof H. Õimu eestvedamisel.

Tallinnas on arvutuslingvistikaalane uurimistöö toimunud põhiliselt Keele ja Kirjanduse Instituudis (praeguses Eesti Keele Instituudis). 1965. aastal loodi Valmen Hallapi algatusel eksperimentaalfoneetika laboratoorium (juhatajaks Georg Liiv, 1971. aastast Arvo Eek), kus tegeldi kõne akustilise ja artikulatoorse uurimisega ning selleks vajaliku tehnilise baasi loomisega. 1977. aastal asutati Mart Remmeli eestvõttel selle laboratooriumi baasil arvutuslingvistika sektor. Olulisemad uurimisteemad on olnud foneetika, sõnastike ja tekstide automaattöötlus, automaatne morfoloogia. Keele ja Kirjanduse Instituut oli omal ajal NSV Liidus esimene, kes võttis sõnaraamatute koostamisel kasutusele arvutid. Ülle Viksi koostatud "Väike vormisõnastik"[10] on olnud aluseks mitme keeletehnoloogiatoote loomisel (sh eesti keele morfoloogiaanalüsaator, õigekirjakontrollija, leksikonid). Küberneetika Instituudiga koostöös valmis kõnesüntesaator, tegeldi ka kõneanalüüsiga. Praegu jätkub kõnetöötlusalane uurimistöö Küberneetika Instituudis Einar Meistri juhendamisel.

Hetkeseis. Tartu Ülikooli arvutuslingvistika töörühmas (*http://www.cl.ut.ee/*) valmis aastail 1991—1996 miljonisõnaline kirjakeele korpus (nn baaskorpus), kuhu on valitud kindlate kriteeriumide alusel eestikeelseid tekste, mis pärinevad aastatest 1983—1987. Korpus on varustatud liidesega, mis võimaldab esitada päringuid, et otsida korpusest teatavatele tingimustele vastavaid lauseid (kasutatav ka veebis). Loodud või loomisel on ka 1970-ndate, 1960-ndate, 1950-ndate, 1930-ndate ja 1890-ndate aastate kirjakeele korpus. Kompleksne eesti kirjakeele korpus moodustab olulise keeleressursi, mida kasutatakse tema väiksusest hoolimata juba praegu mitme keeletöötlusprogrammi koostamisel. On alustatud ka eestikeelse kõnekorpuse tegemist. Alates 1993. aastast on korraldatud seminare, nn korpusepäevi, kus on esinenud nii ülikooli arvutuslingvistid kui ka Eesti teiste teadusasutuste ja välismaa teadlased. Praegu luuakse töörühmas eesti keele semantilist sõnastikku, mille põhimõtteliseks aluseks on WordNet, konkreetseks eeskujuks aga EuroWordNet.

1993. aastal asutati keeletarkvara tootmisele spetsialiseerunud firma Filosoft (*http://www.filosoft.ee/*), mida juhib TÜ üldkeeleteaduse õppetooli teadur Heiki-Jaan Kaalep (seetõttu on firmal ja arvutuslingvistika uurimisgrupil tihedad sidemed). Firma toodeteks on näiteks eesti keele morfoloogia-analüsaator ning sellel põhinevad eesti keele õigekirjakontrollija, poolitaja ja tesaurus, mis ühtlasi kuuluvad paketi MS Office 97 koosseisu. Mitu toodet on kättesaadavad veebis. Arvutuslingvistika uurimisgrupiga teeb koostööd ka TÜ arvutiteaduse instituut, kus praegu tegeldakse eesti keele süntaksianalüüsi automatiseerimise probleemidega.

Tallinnas Eesti Keele Instituudis (*http://www.eki.ee*) on loodud ja luuakse ka veebis kättesaadavaid keeleressursse: sõnastikke, tekstikogusid (1993. aastal alustatud täiendatav ehk avatud kirjakeele korpus, "Väike murdesõnastik" I—II, Asta Õimu fraseoloogiasõnaraamat, sünonüümisõnastik, antonüümisõnastik, Mai Loogi slängisõnaraamat jt), aga ühtlasi ka nende ressursside kasutamise vahendeid (alustades lihtsast sõnastikuotsingust kuni lingvistiliste teisendusteni, mis eeldavad automaatset analüüsi ja sünteesi).

[10] Ü. Viks, Väike vormisõnastik I. Sissejuhatus & grammatika; II. Sõnastik & lisad. Tallinn, 1992.

Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris koostavad E. Meister ja A. Eek eesti keele foneetilist andmebaasi, mida rakendatakse foneetikaalastes uuringutes ja kõnetehnoloogiatoodete väljatöötamisel.

Tartu Ülikool, Eesti Keele Instituut, Küberneetika Instituut ja firma Filosoft on osalenud ja osalevad mitmes rahvusvahelises keeletehnoloogiaalases töös (Euroopa Komisjoni poolt finantseeritavad MULTEXT-EAST, GLOSSER, TELRI, PAROLE-EAST, BABEL, CONCEDE, EuroWordNet-2 jt).

Arvutuslingvistikaalaseid uuringuid on toetanud Eesti Teadusfond (eesti keele korpuse, eesti keele semantilise sõnastiku väljatöötamine, morfoloogiline ühestamine, kõnekorpus) ja Avatud Eesti Fond (projekt Stylus, TÜ arvutuslingvistika kursuste komplekteerimine, projekt KeeleWeb, mille eesmärgiks on süstematiseerida kogu informatsioon, mis Eestis keeletarkvara ja sõnastike kohta olemas on, välja töötada meetodid, mis võimaldavad tarkvara ja sõnastikud ühtseks tervikuks ühendada, ning integreerida see tervik veebi keskkonda avalikuks kasutamiseks).

1997. aastal käivitus Eesti Informaatika Keskuse keeletehnoloogia raamprogramm, kus osalevad TÜ, Filosoft, EKI ja Eesti Õigustõlkekeskus ning mille käigus luuakse eesti keele automaatse analüüsi vahendeid, mida saaks rakendada tekstiandmebaaside koostamiseks ja nendega töötamiseks, sealhulgas intelligentse kasutajaliidese loomiseks.

Eesti arvutuslingvistidel on teadussidemeid ülikoolide ja uurimiskeskustega üle kogu maailma, eriti Helsingi, Stockholmi, Manchesteri, Koblenz-Landau, Zürichi, Praha ja Budapesti ülikooliga ning Pisa Arvutuslingvistika Instituudiga.

Õpetamine. 1997/98. õ-a sügissemestrist alustati Tartu Ülikooli filosoofiateaduskonna eesti filoloogia osakonnas süstemaatilist arvutuslingvistika õpetamist. Seni oli arvutuslingviste ette valmistatud individuaalplaani alusel kas eesti keele või informaatika üliõpilaste hulgast. HESP-i (Higher Education Support Program) ja Avatud Eesti Fondi toetusel töötati eelnevalt välja arvutuslingvistika õppekava, milles on neli aineplokki: keeleteadus, matemaatika, informaatika ja arvutuslingvistika. HESP on toetanud ka mõne arvutuslingvistika kursuse ettevalmistamist ja raamatukogu õppekirjandusega komplekteerimist. Nende kursuste materjalide veebivariandi väljatöötamist toetas Avatud Eesti Fond.

Selles ajakirjanumbris tutvustavad Tartu ülikooli arvutuslingvistid oma tööd. Tinglikult võib artiklid jaotada kolme gruppi: 1) üldülevaated, 2) morfoloogia- ning 3) süntaksi- ja semantikakäsitlused. Autoritest kuuluvad üldkeeleteaduse õppetooli arvutuslingvistika uurimisgruppi Haldur Õim, Kadri Muischnek, Heiki-Jaan Kaalep ja Tarmo Vaino (kaks viimast ka firmasse Filosoft) ning üldkeeleteaduse magistrandid Heili Orav ja Kadri Vider. Mare Koit on arvutiteaduse instituudi dotsent, Tiina Puolakainen ja Kaili Müürisep sama instituudi doktorandid ning Heli Uibo magistrant.