RATSIONALISM JA EMPIRISM KEELE-TÖÖTLUSES: VASTASSEIS VÕI KOOSTÖÖ?*

Mare Koit
Tartu Ülikool

1. Sissejuhatus

Mõni aasta tagasi lahvatas postiloendis CORPORA[1], mille tellijaskond koosneb põhiliselt korpuslingvistidest, diskussioon ratsionaalse (reeglipõhise) ja empiirilise (andmepõhise) keeletöötluse üle. Kõik algas ühe programmeerija pahaaimamatust küsimusest, kas keegi soovitaks tema kiire projekti jaoks statistilist analüüsiprogrammi, kus oleks rakendatud Zellig Harrise transformatsioonilist grammatikat? Teatavasti oli Harris (1909–1992) empiirilise keelekäsitluse pooldaja, aga tema kuulsast õpilasest Noam Chomskyst (sünd 1928) sai hiljem hoopis ratsionalismi eestvõitleja. Diskussiooni käigus tuli korduvalt esile empiirilise ja ratsionaalse keeletöötluse vastasseis kuni selleni, et Chomskyt süüdistati keele automaattöötluse arengu pidurdamises.

2. Ratsionalism ja empirism loomuliku keele automaattöötluses

Ratsionalism väidab, et keelestruktuurid on inimesel kaasa sündinud. See eeldus on aluseks reeglipõhisele keeletöötlusele: keelemudel tuleb arvutile ette anda. Empirism seevastu on seisukohal, et keelestruktuurid õpitakse ainult kogemusest. Seega on keeletöötlus korpustesse koondatud keeleandmete statistiline töötlus ja masinõpe[2], arvuti tuletab keelemudeli ise.

* Autor tänab anonüümseid retsensente, kelle soovitusel on artiklisse lisatud mõistete selgitused.

[1] Postiloend on arhiveeritud aadressil http://listserv.linguistlist.org/archives/corpora.html (8.11.2006).

[2] Automaatne protsess, millega arvuti täiustab oma talitlust, omandades uusi teadmisi ja oskusi või seniseid ümber korraldades.

Loomuliku keele arvutitöötlusest saame rääkida muidugi alles sestpeale, kui ilmusid esimesed elektronarvutid, st alates 1940. aastatest. Huvitav on siinjuures märkida, et arvutite üks esimesi rakendusi oligi seotud keelega (masintõlge).

Kuni 1950. aastate lõpuni valitses keeleteaduses ja -töötluses empirism. Seoses Chomsky silmapaistvate artiklitega algas aga 1960. aastatel ratsionalismi võidukäik, mis kestis peaaegu 1980. aastate lõpuni. Sellest ajast alates kuni tänaseni on jälle domineerinud empirism.

2.1. Empirismi algus

Empiirilise keeletöötluse alusepanijaks võib lugeda vene matemaatikut Andrei Markovit (1856–1922), kes oma 1913. a avaldatud artiklis kasutas Markovi ahelaid[3] ennustamaks, kas järgnev täht romaanis „Jevgeni Onegin" on vokaal või konsonant, arvestades ühte või kahte eelnevat tähte, st kasutades bigramm- ja trigramm-mudeleid.

1947. a püstitas ameerika matemaatik Warren Waver kirjas küberneetika rajajale Norbert Wienerile küsimuse, kas arvuteid saaks kasutada tekstide tõlkimisel ühest keelest teise, ja soovitas käsitleda masintõlget kui salakirja dešifreerimist (avaldati memorandumina 1949). See idee leidis maailmas suurt vastukaja ja käivitas aktiivse masintõlkealase tegevuse, mis 1954. a jõudis firma IBM esimese avaliku eksperimendini: arvuti tõlkis 200-sõnalise teksti vene keelest inglise keelde. Tõsi, enamasti piirduti tol ajal sõna-sõnalise tõlkimisega ega saavutatud seetõttu kuigi head kvaliteeti. Muidugi oli takistuseks ka tagasihoidlik arvutustehnika.

1949. a genereeris informatsiooniteooria[4] rajaja Claude Shannon inglise keele n-gramm-sõnamudelid[5] 1950. aastate strukturaalse lingvistika põhimõisteks sai korpus – teataval viisil kogutud ja süstematiseeritud keeleandmete kogu.

1957 a esitas Frank Rosenblatt lihtsa närvivõrgu matemaatilise mudeli – tajuri (ingl *perceptron*), luues seega eeldused konnektsio-

---

[3] Statistiline mudel, mis koosneb sõlmedest (olekutest) ja neid ühendavatest suunatud kaartest (üleminekutest), kus iga kaar on varustatud tema läbimise tõenäosusega.

[4] Informatsiooni kvantitatiivsete mõõtudega tegelev teadusharu.

[5] Statistiline keelemudel, kus sõna esinemise tõenäosus sõltub n-1 talle eelnevast sõnast.

nistlike keelemudelite[6] kasutuselevõtuks ja keeleandmete paralleeltöötluseks.

1960. aastatel loodi esimesed arvutikorpused: Browni, London-Lundi, Lancaster-Oslo-Bergeni korpus – igaühes miljon sõnet. Sestpeale sai empiiriline materjal statistiliste keelemudelite treenimiseks elektroonilisel kujul kättesaadavaks. Töötati välja Markovi ahela edasiarendus – Markovi peitmudel[7] (*Hidden Markov Model*, HMM). 1967. a rakendati antud sõnajärjendi jaoks kõige tõenäolisema sõnaliikide järjendi leidmiseks Viterbi algoritmi[8] Loodi esimesed tõenäosuslikud grammatikad[9]

Kuid vahepeal oli esile kerkinud ratsionalistlik keelekäsitlus, seda eeskätt Chomsky tööde tõttu.

2.2. Ratsionalismi algus: Chomsky generatiivsed grammatikad

Oma artiklis (Chomsky 1956) defineeris Chomsky generatiivse grammatika, mis sai keelte automaattöötlusel kasutatavate formalismide ja mudelite põhiliseks aluseks ligi kolmekümne järgneva aasta vältel.

Generatiivne grammatika on nelik G = (**T**, **N**, **P**, S), kus **T** on terminaalide (nt keele sõnavormide), **N** mitteterminaalide (nt grammatiliste kategooriate) ja **P** ümberkirjutusreeglite hulk kujul x → y (tähenduses: kirjuta x asemele y; siin on x ja y terminaalide ja/või mitteterminaalide järjendid, nn sõned, kusjuures x ei ole tühisõne) ning S üks erilises rollis mitteterminaal – grammatika lähtesümbol (*Sentence* 'lause'). Kõikvõimalike terminaalsete järjendite hulka, mida saab sümbolist S lähtudes ja ümberkirjutusreegleid rakendades tuletada, nimetatakse selle grammatikaga genereeritavaks keeleks.

Joonisel 1 on esitatud grammatika, mis genereerib väikese osa eesti keelest: laused *Mari kirjutab*, *Jüri kirjutab*, *Väike Mari loeb hästi*, *Väike tubli Jüri kirjutab hästi meelsasti* jms.

---

[6] Matemaatiline mudel, mis esitatakse kui teatavate lihtsate komponentide võrgustik.

[7] Statistiline mudel, kus vaadeldavate olekute alusel määratakse peitolekud (nt sõnade järjendi alusel sõnaliigid).

[8] Algoritm Markovi peitmudelis vaadeldavate olekute järjendi alusel kõige tõenäolisema peitolekute järjendi määramiseks.

[9] Grammatika, kus iga reegel on varustatud tema rakendamise tõenäosusega.

G = (**T**, **N**, **P**, S), kus
T = {*väike:*Adj, *tubli:*Adj, *Mari:*N, *Jüri:*N, *kirjutab:*V, *laulab:*V, *hästi:*Adv, *meelsasti:*Adv}
eesti keele sõna(vormi)d
N = {S, VP, NP, N, V, Adj, Adv}
grammatiliste kategooriate nimetused
P = {S → NP VP, NP → Adj NP, NP → N, VP → VP Adv, VP → V}
ümberkirjutusreeglid

Joonis 1. Generatiivne grammatika eesti keele väikese osa jaoks

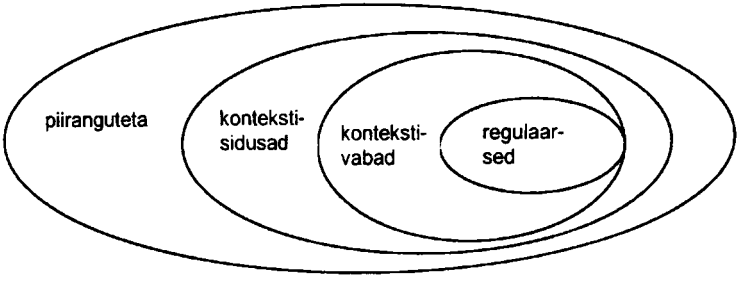
Generatiivsed grammatikad ja nendega genereeritavad keeled moodustavad nn Chomsky hierarhia, sõltuvalt ümberkirjutusreeglite kujust (vt tabel 1 ja joonis 2). Kõige üldisemad on 0-tüüpi e piiramata grammatikad, kus reeglite kujule pole seatud täiendavaid kitsendusi. 1. tüüpi e kontekstisidusates grammatikates esineb reegli vasakul ja paremal pool üks ja sama kontekst. 2. tüüpi e kontekstivabades grammatikates on reegli vasakul pool ainult mitteterminaal. 3. tüüpi e regulaarses grammatikas esineb reegli vasakul pool üks mitteterminaal nagu kontekstivabas grammatikas, kuid paremal pool tohib olla kas üks või kaks sümbolit.

Tabel 1. Ümberkirjutusreeglite kuju Chomsky hierarhias

Grammatika / keele tüüp	**Ümberkirjutusreeglite kuju**
0 (piiranguteta)	$x \rightarrow y$, kus $x \in (\mathbf{T} \cup \mathbf{N})^+$ $y \in (\mathbf{T} \cup \mathbf{N})^*$
1 (kontekstisidus, ingl *context sensitive*)	$zAw \rightarrow ztw$, kus $A \in \mathbf{N}$, $t \in (\mathbf{T} \cup \mathbf{N})^+$ $z, w \in (\mathbf{T} \cup \mathbf{N})^*$
2 (kontekstivaba, *context free*)	$A \rightarrow y$, kus $A \in \mathbf{N}$, $y \in (\mathbf{T} \cup \mathbf{N})^*$
3 (regulaarne, *regular*)	$A \rightarrow a$, $A \rightarrow Ba$ (vasaklineaarne), kus $A, B \in \mathbf{N}$, $a \in \mathbf{T}$

Chomsky tõestas, et loomulike keelte süntaks ei ole kirjeldatav regulaarse grammatikaga, ja püstitas hüpoteesi, et loomulikud keeled on koguni kontekstisidusad (tõsi küll, märkides, et nähtused, mis arvatavasti ei ole kirjeldatavad kontekstivabade reeglitega, on keeles suhteliselt harvad).

Chomsky tööd pälvisid kogu maailmas suurt tähelepanu. Paljude keelte jaoks hakati looma generatiivseid grammatikaid. Püüti matemaatiliselt põhjendada Chomsky hüpoteesi. 1985. a tõestati, et Šveitsi saksa keel ei ole kontekstivaba (Schieber 1985).



Joonis 2. Chomsky grammatikate ja keelte hierarhia

Ka Tartu ülikoolis alustas 1960. aastate keskpaigas keeleteadlase Huno Rätsepa juhtimisel tööd seminar (hilisema nimetusega Generatiivse Grammatika Grupp e GGG), kus keele arvutitöötlusest huvitatud lingvistika- ja matemaatikaüliõpilased ning -õppejõud uurisid Chomsky artikleid ja arutasid eesti keele generatiivse grammatika loomise üle. Veelgi varem, 1950. aastate lõpus, oli matemaatikaõppejõud Ülo Kaasik algatanud koos üliõpilastega vene-eesti masintõlkeprojekti, kus töötati välja venekeelse matemaatilise teksti morfoloogilise analüüsi reeglid ja koostati programm arvutile Ural (Kaasik, Korjus 1959; Palm 1962).

2.3. Chomsky argumendid empirismi vastu

Chomsky tegi vahet keelepädevuse ja -kasutuse vahel. Keelepädevus on meie sisemine teadmine keelest. Ta selgitab ja iseloomustab rääkija keeleteadmust. Keelekasutus on üksnes keelepädevuse peegeldus, selle väline avaldus, millele avaldavad mõju mitmed keelevälised tegurid.

Chomsky arvates peab arvutis modelleerima just keelepädevust, aga mitte keelekasutust, sest kui me ei suudaks modelleerida keelepädevust, siis kuidas me oskaksime suvalise väljendi kohta otsustada, kas see on korrektne keelekasutus.

Keelepädevust kasutamata, ainuüksi korpuse põhjal, ei suudaks me näiteks otsustada, et järgnevast loetelust esimene lause pole korrektne, aga ülejäänud on:

**Ta paistab Tõnule raamatud.*
Ta annab Tõnule raamatud.
Ta laenab Tõnule raamatud.
Ta võlgneb Tõnule raamatud.

Chomsky väitis, et korpus on oma olemuselt keelekasutuse väljendus ega sobi seetõttu keelepädevuse modelleerimiseks. Ta kritiseeris varase korpuslingvistika lähteseisukohti (loomuliku keele laused on lõpliku pikkusega ning neid saab koguda ja loendada), väites vastu, et lausete hulk on potentsiaalselt lõpmatu ja laused võivad seejuures olla kui tahes pikad, nt *The cat (the dog (the rat (...) bit) chased) died.* Sellest tegi Chomsky järelduse, et keele grammatika koostamiseks ei tule koguda korpust, vaid kirjeldada keele reeglid (mida on lõplik hulk).

Chomsky kriitika määras keele automaattöötluse arengusuuna ligi kolmekümneks aastaks.

2.4. Ratsionalismi domineerimine

Chomsky generatiivsetest grammatikatest inspireerituna töötati välja ja võeti keeletöötluses kasutusele veel suur hulk mitmesuguseid reeglipõhiseid formalisme, sh transformatsioonigrammatika (1965, Chomsky), käändegrammatika (*case grammar*, 1967, Ch. Fillmore), laiendatud üleminekuvõrk (*augmented transition network* e ATN, 1970, W Woods),

määravate osalausete grammatika (*definite clause grammar* e DCG, 1978, A. Colmerauer) jpm.

Alates 1980. aastatest hakati looma ja rakendama mitmesuguseid nn unifikatsioonigrammatikaid, kus reeglid ei opereeri ainult süntaktiliste kategooriatega, vaid kui tahes keeruliste tunnusestruktuuridega. Tunnused võivad olla nii morfoloogilised, süntaktilised kui ka semantilised. Tuntud formalismid on näiteks funktsionaalne unifikatsioonigrammatika (*Functional Unification Grammar* e FUG, 1979, M. Kay), leksikaalfunktsionaalne grammatika (*Lexical Functional Grammar* e LFG, 1982, J. Bresnan ja R. Kaplan), üldistatud fraasistruktuurigrammatika (*Generalized Phrase Structure Grammar* e GPSG, 1985, G. Gazdar) ning peajuhitav fraasistruktuurigrammatika (*Head driven Phrase Structure Grammar* e HPSG, 1987, C. Pollard ja I. Sag). Neist viimast on kasutatud nt inglise, saksa, vene, tšehhi ja bulgaaria keele analüsaatorite ning generaatorite loomisel[10]

Kõrvuti grammatikatega said populaarseks reeglipõhiseks formalismiks ka olekuautomaadid[11], millest lihtsaim on lõplik automaat. Lõplikku automaati võib esitada nn olekudiagrammina. Lause (või sõna) analüüs tähendab siis diagrammi läbimist algolekust lõppolekuni, liikudes mööda olekuid ühendavaid kaari, mis on märgendatud lauses esinevate sõnavormidega (või vastavalt sõnas esinevate morfeemidega). Lõplikud automaadid on leidnud rakendust eeskätt morfoloogiamudelina. Hästi tuntud kahetasemelist morfoloogiamudelit[12] on kasutatud paljude keelte, sh ka eesti keele automaatseks morfoloogiliseks analüüsiks ja genereerimiseks (Uibo 2006).

2.5. Empirismi areng

1967. a näitas E. M. Gold, et korrektset kontekstivaba grammatikat pole võimalik usaldusväärselt õppida selle grammatika lausetest ehk ainult positiivsetest näidetest[13] Prominentsed lingvistid eesotsas Chomskyga

[10] http://www.ling.ohio-state.edu/research/hpsg/ (8.11.2006)

[11] Matemaatiline mudel, mis koosneb sõlmedest (olekutest) ja neid ühendavatest suunatud kaartest (üleminekutest).

[12] Mudel eristab süva- ja pindmist taset: sõnastikus säilitatakse morfeemide nn süvakujusid, millest reeglite ja sõnastikevaheliste viitade abil saab moodustada kõik tegelikkuses esinevad sõnavormid.

[13] http://www.isrl.uiuc.edu/~amag/langev/paper/gold67limit.html (8.11.2006)

kasutasid seda tulemust kui tõendust kaasasündinud universaalse grammatika olemasolule. Chomsky väitis, et lastel, kes alles hakkavad keelt omandama, pole negatiivseid keelenäiteid, sest nende vanemad ja hoidjad produtseerivad enamasti õigeid lauseid ja parandavad vigu väga harva (seda nimetas ta stiimuli viletsuseks (*poverty of the stimulus*)). Järelikult peavad lapsed juba sünnist saadik teadma grammatikat ning keele omandamine tähendab lihtsalt grammatika mõne parameetri häälestamist ja sõnavara õppimist.

Empirismi poolehoidjatele andis aga tuge 1969. a tõestatud tulemus (Horning 1969), et **tõenäosuslikke** grammatikaid **saab** õppida ainult positiivsetest näidetest (see on nn tõenäoliselt ligikaudu korrektne õppimine, ingl *probably approximately correct*). Tõenäosuslikus grammatikas on iga reegel varustatud tema rakendamise tõenäosusega, mis on arvutatud korpuse põhjal. Ka olekuautomaadile võib lisada ühest olekust teise üleminekute tõenäosused – nii saame Markovi mudeli.

1976. aastal hakati tõenäosuslikke mudeleid kasutama kõnetuvastuses. 1985. a õnnestus K. Churchil kõne sünteesimisel edukalt määrata pärisnimede päritolu, kasutades statistilisi trigramm-mudeleid. Sestpeale muutusid statistilised keeletöötlusmeetodid järjest populaarsemaks ja 1990. aastatest alustasid nad oma võidukäiku, tõrjudes kõrvale reeglipõhised meetodid. Miks? Statistilised mudelid suudavad toimida hästi ka siis, kui teadmus on mittetäielik, mistõttu neid ongi palju ja edukalt rakendatud just kõnetuvastuses ja -sünteesis.

Kõrvuti statistilistega on tuntud veel teinegi liik andmepõhiseid keelemudeleid – konnektsionistlikud, nt tehisnärvivõrgud. Konnektsionistlik mudel koosneb suurest hulgast omavahel seotud lihtsatest mittelineaarsetest komponentidest. Komponendid töötavad paralleelselt (mitte järjestikku nagu lõplik automaat). Siin kasutatakse samuti nagu statistilise mudeli puhul treeningandmeid, mille põhjal mudel n-ö õpib, kuid keeletöötlussüsteemi arhitektuur on keerukam ja tänu sõlmede dubleerimisele on tehisnärvivõrk töökindlam.

2.6. Empirism ratsionalismi vastu

Kui ratsionalistid kirjeldavad inimajus asuvat keelemudelit (nn I-keelt, ingl *internal*), siis tegelikud keeleandmed (nn E-keel, ingl *external*) on üksnes kaudne tõendusmaterjal I-keele kohta. I-keel viitab indiviidi aju sisemisele seisundile. See, et inimesed räägivad ühes ja samas keeles,

eeldab, et neil on enam-vähem üks ja sama I-keel. Antud lausete hulga (E-keele) kirjeldamiseks võib aga edukalt kasutada erinevaid grammatikaid. Empirismi põhiargument ratsionalismi vastu ongi see, et keelepädevust pole võimalik isoleeritult käsitleda – kirjeldada saab üksnes tegelikku keelekasutust.

Lähtudes tõsiasjast, et inimene tuvastab paremini sagedasemaid sõnu ja konstruktsioone, tuleks ka keeletöötlusel keskenduda tavalisele, sagedasele.

Reeglipõhine keeletöötlus oli asjakohane, kuni piirduti väikeste („mängu-") süsteemide loomisega (muidugi oli see tingitud ka arvutustehnika võimalustest).

3. Head ja vead

Nii reegli- kui ka andmepõhisel keeletöötlusel on oma positiivsed ja negatiivsed küljed. Reeglipõhiste keelemudelite põhiliseks eeliseks on arusaadavus kirjeldava ja genereeriva jõu mõttes ning praktilistes rakendustes, sest mudeli loonud inimene on oma keeleteadmusest just sellised reeglid tuletanud. Reeglipõhised mudelid suudavad aga efektiivsemalt käsitleda kaugsõltuvusi (nt aluse ja öeldise ühildumist), on läbinähtavad (st lingvistilised faktid on mudeli struktuuris ja koostisosades selgelt väljendatud) ning pööratavad (st rakendatavad nii analüüsiks kui ka genereerimiseks).

Samas on reeglipõhistel keelemudelitel kaalukad puudused: nad on haprad praktilistes rakendustes ning tundlikud sisendi väikestegi kõrvalekallete ja ebaregulaarsuste suhtes (mis teeb nad sobimatuks näiteks kõnetuvastuses). Mudelite väljatöötamiseks on vaja häid eksperte, kuna sellised mudelid ei suuda ise näidetest õppida, samuti on nende modifitseerimine raske.

Andmepõhiste keelemudelite põhiline eelis on see, et kui neid on treenitud korpustel, siis suudavad nad efektiivselt käsitleda tüüpilist keelekasutust. Nad ületavad reeglipõhiseid mudeleid selliste lingvistiliste nähtuste modelleerimisel, mille kohta inimestel pole veel selget arusaama, nt kõne. Mudeli efektiivsus sõltub oluliselt treeningandmete mahust: suurem andmete hulk on parem. Andmepõhise keeletöötluse eelis on ka see, et tulemusi saab kiiresti.

Andmepõhise keeletöötluse oluline puudus on asjaolu, et kaob lingvistile vajalik keelemudeli intuitiivne selgus ja hoomatavus. Programm

teeb midagi ja teeb seda hästi, aga miks – see pole päris selge. Andmepõhiste keelemudelite teine puudus on treenimiseks vajalike korpuste kogumise ja märgendamise töömahukus ja veaohtlikkus. Mudeli täitmisomadused sõltuvad erinevate klasside arvust: mida rohkem klasse (nt erinevaid sõnaliike), seda raskem on nii treenimine kui ka juba treenitud süsteemi töö.

4. Koostöö

Millist meetodit eelistada? Konkreetse meetodi valik sõltub esmajoones valdkonnast ja rakendusest. Kõnetöötluses, kus on tegu andmete suure variatiivsusega, sobivad paremini empiirilised meetodid. Need on orienteeritud statistilisele keskmisele ja suudavad edukalt käsitleda sagedasi nähtusi, küll aga mitte kõrvalekaldeid keskmisest. Selleks, et automaatselt tuvastada nt kõnepuudelise kasutaja kõnet, oleks vaja koguda erikorpusi.

Morfoloogilisel ja süntaktilisel analüüsil on eelistatud reeglipõhiseid meetodeid, sest siin saab välja tuua kindlapiirilised normid ja reeglid.

Ka eesti keele automaattöötlusel on kasutatud erinevaid meetodeid (vt tabel 2). Kõnetuvastuses ja -sünteesis, millega tegeldakse Tallinna Tehnikaülikooli Küberneetika Instituudi foneetika ja kõnetehnoloogia laboris ja Eesti Keele Instituudis, kasutatakse andmepõhiseid meetodeid (Alumäe 2005; Mihkla jt 1999). Morfoloogiatarkvara loomisel on nii Tartu Ülikoolis, Eesti Keele Instituudis kui ka keeletarkvarafirmas Filosoft kasutatud reeglipõhiseid meetodeid (Uibo 2006; Kaalep, Vaino 2000; Viks 2000). Morfoloogilisel ühestamisel on rakendatud nii andmepõhist kui ka reeglipõhist meetodit (Kaalep, Vaino 1998; Roosmaa jt 2003). Eesti keele süntaksianalüsaator on reeglipõhine (Roosmaa jt 2003). Sõnatähenduste ühestaja kombineerib tesauruses reeglistatud maailmateadmust ja masinõpet (Vider, Kaljurand 2002). Dialoogiaktide tuvastamisel on katsetatud andmepõhiseid meetodeid (Fišel, Kikas 2006).

Tuleb rõhutada, et ka andmepõhine keeletöötlus ei saa alata tühjalt kohalt. Programme treenitakse märgendatud korpustel, aga korpuste märgendamisel on alati aluseks võetud mingi formalism, reeglistik – seega on inimene programmile juba ette andnud teatava keeleteadmuse. Näiteks on eelnevalt kindlaks määratud sõnaliikide nimestik või on

korpus süntaktiliselt analüüsitud, kasutades teatavat fraasistruktuurigrammatikat.

Sageli kombineeritakse erinevaid meetodeid sel viisil, et ühel keele analüüsimise või sünteesimise etapil kasutatakse andmepõhist, teisel aga reeglipõhist meetodit. Näiteks dialoogsüsteem SCREEN tuvastab kõnet, kasutades selleks statistilist ja konnektsionistlikku mudelit, ning seejärel toimub reeglipõhine süntaktiline, semantiline ja dialoogi analüüs; SUITE tuvastab kõnet, kasutades konnektsionistlikku mudelit foneemide identifitseerimiseks, statistilisi meetodeid parima lausehüpoteesi loomiseks ning reeglipõhist meetodit täiendavaks lingvistiliseks analüüsiks (Manaris 1998).

Hübriidsed tehnikad vähendavad inimese jõupingutusi keelemudeli konstrueerimisel ning tõstavad samas süsteemi paindlikkust, efektiivsust ja tõrkekindlust.

Tabel 2. Eesti keele automaattöötlusel kasutatud meetodid

Keeletöötluse tase	*Kus tegeldakse*	*Meetodid*
SUULINE KEEL	TTÜ Kübl + EKI	andmepõhised
kõne => tekst,		
tekst => kõne		
KIRJALIK KEEL (tekst)		
Morfoloogia	Filosoft + TÜ	reeglipõhine + andmepõhine
	EKI	reeglipõhine
		(avatud morfoloogiamudel)
	TÜ	reeglipõhine (lõplikud automaadid)
Süntaks	TÜ	reeglipõhine
		(kitsenduste grammatika)
Semantika		
sõnatähenduste	TÜ	andmepõhine + reeglipõhine
ühestamine		
Pragmaatika	TÜ	
dialoogiaktide		andmepõhised
tuvastamine		(tehisnärvivõrgud, otsustuspuud[14])

Mitmel hiljutisel rahvusvahelisel keeletehnoloogia konverentsil on tõstatatud küsimus paradigma vahetuse vajalikkusest. Sheffieldi ülikooli professor Roger Moore märkis konverentsi „Kõne ja arvuti" (SPECOM)

[14] Hierarhiliste otsuste graafiline esitus, mis viib minimaalse arvu otsuseni.

plenaarettekandes 2005. a, et kuigi 50 aasta jooksul on kahtlemata toimunud suur progress meie teaduslikus arusaamises sellest, kuidas inimesed kasutavad ja töötlevad suulist keelt, ning meie tehnilises võimes jäljendada sellist käitumist praktilistes arvutisüsteemides, pole see paraku tingitud sellest, nagu saaksime nüüd palju paremini aru inimese keeletöötlusprotsessidest, vaid pigem tänu masinõppele ja arvutite jõudluse kasvule. Ta rõhutas, et on vaja paradigma vahetust algoritmides ja tehnoloogias – unifitseeritud teooriat, mis ühendaks erinevad keeleuurimisega tegelevad teadusharud: lingvistika, arvuti- ja psühholingvistika, tehisintellekti, psühholoogia jmt. Moore nimetas uut teooriat kognitiivseks informaatikaks (Moore 2005, vt ka Einar Meistri artiklit käesolevas kogumikus).

5. Kokkuvõte

Loomuliku keele automaattöötluses on erinevatel perioodidel domineerinud erinevad lähenemisviisid. **Ratsionalistliku** keelekäsitluse kohaselt, mille tõi keeletöötlusse 1950. aastate lõpul Noam Chomsky oma generatiivse lingvistikaga, tuleb modelleerida keelepädevust – inimese sisemist teadmust keele kohta. Keele tegelik kasutus on üksnes keelepädevuse ilming ja seda võivad mõjutada mitmesugused keelevälised tegurid, sh rääkija emotsionaalne seisund või ümbritsev müra. Chomsky väitis, et loomuliku keele lausete hulk on potentsiaalselt lõpmatu ja seega ei saa keele grammatika koostamiseks lihtsalt loendada lauseid, vaid tuleb kirjeldada keele reeglid (mida on lõplik hulk). Usuti, et märkimisväärne osa teadmistest inimajus on ette fikseeritud, ehk isegi geneetilise päritoluga.

Ratsionalism vastandus **empirismile**, mida oli juba Chomsky-eelsel (ja arvutite-eelsel) ajal arendanud strukturaalne lingvistika. Empirism eeldab samuti mõningate kognitiivsete võimete olemasolu: arvestades, et õppimine pole võimalik täiesti puhtalt lehelt, peab ajus eksisteerima teatav algstruktuur, mis seejärel tagab meelte kaudu saadud andmete organiseerimise ja üldistamise. Loomuliku keele automaattöötluseks tuleks siis kõigepealt luua üldine keelemudel, misjärel toimub mudeli parameetrite väärtuste tuletamine sel teel, et suurel keeleandmete hulgal (keelekorpustel ja -andmebaasidel) rakendatakse statistilisi meetodeid, kujundite tuvastamist ning masinõpet.

Tänu andmepõhiste meetodite kasutuselevõtule on saavutatud keeletöötluses suurt edu. Ometi on näiteks kas või kõnetuvastuse kvaliteet seni veel mitu korda kehvem kui inimesel. Aeg on küps uue, kõiki inimkeele uurimise ja modelleerimisega seotud valdkondi ühendava teooria loomiseks.

Kirjandus

Alumäe, Tanel 2005. Using Adaptive Stochastic Morphosyntactic Language Model for Two-pass Large Vocabulary Estonian Speech Recognition. – Proceedings of SPECOM, 10th International Workshop on Speech and Computer. Toim. G. Kokkinakis, N. Fakotakis, E. Dermatas, R. Potapova. Patras, 515–518.

Chomsky, Noam 1956. Three models for the description of language. IRE Trans. Inf. Th., IT-2.

Fišel, Mark, Taavet Kikas 2006. Dialoogiaktide automaatne tuvastamine. – Keel ja arvuti. (TÜ üldkeeleteaduse õppetooli toimetised 6.) Toim. M. Koit, R. Pajusalu, H. Õim. Tartu, 233–245.

Horning, James Jay 1969. A Study of Grammatical Inference. Ph. D. Dissertation. Computer Science Department, Stanford University, California.

Kaalep, Heiki-Jaan, Tarmo Vaino 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Arvutuslingvistikalt inimesele. (TÜ üldkeeleteaduse õppetooli toimetised 1.) Toim. T. Hennoste. Tartu, 87–99.

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

Kaasik, Ülo, Ain Korjus 1959. Automaatsest tõlkimisest. – Keel ja Kirjandus 11, 663–673.

Manaris, Bill 1998. Natural Language Processing: A Human–Computer Interaction Perspective. – Advances in Computers 47. Toim. M. V Zelkowitz. New York: Academic Press, 1–66. http://www.cs.cofc.edu/~manaris/publications/ advances-in-computers-vol-47.pdf (8.11.2006)

Meister, Einar 2006. Teooria ja praktika vahekorrast kõnetehnoloogias: täiuslikuma tehnoloogia suunas. Käesolevas kogumikus.

Mihkla, Meelis, Arvo Eek, Einar Meister 1999. Text-to-Speech Synthesis of Estonian. – Proceedings of the 6th European Conference on Speech Communication and Technology 5. Budapest, 2095–2098.

Moore, Roger 2005. Cognitive Informatics: The Future of Spoken Language processing? – Proceedings of SPECOM, 10th International Workshop on Speech and Computer. Toim. G. Kokkinakis, N. Fakotakis, E. Dermatas, R. Potapova. Patras, 11–15.

Palm 1962 = Пальм, Рээдик 1962. О морфологическом анализе русской фразы. – Сообщения по машинному переводу 1. Таллинн, 59–83.

Roosmaa, Tiit, Mare Koit, Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Heli Uibo 2003. Eesti keele arvutigrammatika: mis on tehtud ja kuidas edasi? – Keel ja Kirjandus 3, 192–209.

Shieber, Stuart M. 1985. Evidence against the context-freeness of natural language. – Linguistics and Philosophy 8, 333–343.

Uibo, Heli 2006. Eesti keele morfoloogia modelleerimisest lõplike muundurite abil. – Keel ja arvuti. (TÜ üldkeeleteaduse õppetooli toimetised 6.) Toim. M. Koit, R. Pajusalu, H. Õim. Tartu, 13–35.

Vider, Kadri, Kaarel Kaljurand 2002. Automatic WSD: Does it make sense of Estonian? – Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems, 159–162.

Viks, Ülle 2000. Eesti keele avatud morfoloogiamudel. – Arvutuslingvistikalt inimesele. (TÜ üldkeeleteaduse õppetooli toimetised 1.) Toim. T. Hennoste. Tartu, 9–36.