33

MOSAIIK

Haldur Õim

INIMENE, KEEL JA ARVUTI

Haldur Õim

INIMENE, KEEL JA ARVUTI

ehk kompuuterlingvistika

Toimetanud Ustus Agur

Fr. R. Kreutzwaldi nim. ENSV Riiklik Raamatukogu
286.413

TALLINN · «VALGUS» 1983

4
Õ35

Sarja ühiskondlik toimetuskolleegium:
A. Koppel, T. Orav, J. Peegel, J. Selirand

Retsenseerinud M. Hint ja H. Rätsep

Kaane kujundanud T. Aru

Õim, H.

Õ35 **Inimene, keel ja arvuti ehk kompuuterlingvistika.** — Tallinn: Valgus, 1983. — 144 lk., ill. — (Sari «Mosaiik»)

Raamat annab ülevaate keele automaatse analüüsi meetoditest, mida rakendatakse võimaldamaks inimesel ja arvutil suhelda loomulikus keeles. Lähemalt käsitletakse arvutite abil toimuva informatsiooniotsingu, masintõlke ning keelt mõistvate süsteemide olemust, talitlust ja tänapäeva taset.

Populaarne raamat kõigile asjasthuvitatuile.

Õ 1503000000—039
M902(16)—83 18—82

81.2

© Kirjastus «Valgus» 1983

Saateks

Aeg-ajalt on püütud inimkeelt õpetada mõnele mitteinimlikule olendile, näiteks inimahvile. Seda mitte tõsimeelsest lootusest teda tõepoolest kõnelema õpetada või sellest õpetamisest otsest kasu saada, vaid rohkem uudishimust aetuna: mida ja kui palju ta inimeste keelest suudab ära õppida.

Juba alates 1950-ndatest aastatest püütakse keelt õpetada ka elektronarvutile. Ja siin ei saa öelda, nagu ei võetaks asja tõsiselt. On rohkesti neid, kes on veendunud, et kunagi mitte väga kauges tulevikus hakkavad inimesed tõepoolest arvutitega suhtlema loomulikus keeles. Pealegi ei innusta neid jõupingutusi ainult uudishimu (ehkki sellelgi pole siin väike osa), vaid tagant sunnib ka üha tungivam vajadus niisuguse suhtlemisviisi järele.

Vestlusest arvutiga võib praegu rääkida ainult kujundlikus mõttes. Kuid arvutid suudavad keelest juba sel määral aru saada ja tekstidega opereerida, et sellest on põhjust põhjalikumalt rääkida. Käesolev raamat ongi pühendatud peamistele aladele, kus arvutid on kasutatavad loomuliku keele analüüsimiseks. Teisiti öeldes, raamat on pühendatud kompuuterlingvistikale.

Kompuuterlingvistika ehk arvutilingvistika on hübriidteadusharu, mis on sündinud keeleteaduse kui teoreetilise teaduse ja elektronarvutite kasutamise kui praktilise tegevusala piirimail. Sellisena on kompuuterlingvistikal nii omad praktilised ülesanded kui ka teoreetilised eesmärgid. Meie käsitlus liigubki praktilisemate alade juurest teoreetilisemate küsimuste suunas. Räägime automatiseeritud infootsingust, masintõlkest ja tekstide sisu mõistmise modelleerimisest arvutil. Enne konkreetsete probleemide juurde asumist püüame aga esimeses peatükis selgitada üldist tausta kultuuri, teaduse ja tehnika arengu näol, mis on muutnud teksti töötlemise elektronarvutil nii aktuaalseks.

Kõigis peatükkides on probleemidele lähenetud keele, mitte tehniliste realisatsioonide seisukohalt. See tähendab, raamat on mõeldud eelkõige neile, kes on huvitatud rohkem keelest kui arvutitest. Arvuteist otseselt raamatus juttu ei tule. Sageli on populaarteaduslikes raamatutes, milles räägitakse arvutite rakendamisest mõnel uuel alal, muu jutu ees peatükk, milles selgitatakse elektronarvuti ehitust ja töötamispõhimõtteid. Me usume, et käesolevas raamatus pole seda vaja. Esiteks tundub meile, et arvutid on saanud tänapäeva kultuuri lahutamatuks osaks ning igal inimesel on neist juba mingi ettekujutus, mida kümmekonna lehekülje tehniliste kirjeldustega võib ainult segasemaks ajada. Teiseks aga ei peaks järgnevates peatükkides käsitletavate probleemide arusaadavus kuigivõrd oluliselt sõltuma teadmisest, kuidas töötab elektronarvuti.

I. Inimene ja informatsioon

Kus loodus lõpetab oma asjade loomise, seal asub inimene looduslikest asjadest sellesama looduse abil looma uute asjade loendamatuid liike.

Leonardo da Vinci

I.1. Inimene ja informatsioon — poolfilosoofiline sissejuhatus

Informatsioonil on täita oluline ja aina suurenev osa inimeste teaduslikus, majanduslikus ja igapäevases elus. 20. sajandi ühiskonnale kujutab informatsioon endast kümneid kordi suuremat probleemi kui näiteks 10. sajandi ühiskonnale.

Me teame, et ühiskonna eksisteerimise aluseks on majandus — materiaalsete väärtuste tootmine. Selleks et rahuldada inimkonna suurenevaid vajadusi, on vaja toota üha rohkem, paremini ja tõhusamalt, opereerida aina suuremate massi- ja energiakogustega. Et sellega toime tulla, et juhtida neid üha suuremaid mastaape omandavaid protsesse, et välja töötada üha täiuslikumaid tehnoloogiaid, leiutada üha paremaid aparaate, aga lõpuks ka selleks, et paremini tundma õppida iseennast ja oma vajadusi — kõigeks selleks on vaja üha rohkem i n f o r m a t s i o o n i. Selleks et toota kaupu või energiat ning toime tulla ühiskonna juhtimise ja korraldamisega, on juba tänapäeval — aga veelgi rohkem tulevikus — vaja sõna otseses mõttes t o o t a t e a d m i s i, toota neid sihikindlalt ning plaanipäraselt.

Seejuures ei ole majanduse ja muude ühiskondlike tegevusalade areng ühelt poolt ja selle tegevuse juhtimiseks ning arendamiseks vajamineva informatsiooni kasv teiselt poolt võrdelised: mida kõrgemale arengutasemele tõuseb ühiskond ja mida komplitseeritumaks muutuvad selles kulgevad protsessid, seda suurema osakaalu ühiskonna elus omandab informatsioon.

1955. aastal kulutati veerand Ameerika Ühendriikide rahvuslikust koguproduktist «teadmiste tööstusele», s. o. ideede ja informatsiooni ning mittemateriaalsete väärtuste tootmisele. See oli juba kolm korda enam kui Ühendriigid kulutasid «teadmiste sektorile» 1900. aastal. Ent kümme aastat hiljem, 1965. aastal, neelas teadmiste sektor juba ühe kolmandiku vahepeal veelgi

kasvanud rahvuslikust koguproduktist. 1970-ndate aastate lõpuks prognoositi selle sektori arvele juba pool rahvuslikust koguproduktist. «Iga teine dollar, mis teenitakse ja kulutatakse Ameerika majanduses, kulub ideede ning informatsiooni loomiseks ja levitamiseks.»[1]

Ideede ja informatsiooni tootmine ning levitamine kujuneb üha selgemalt sõlmprobleemiks, millest sõltub ühiskonna elujõud ja arengupotentsiaal. Nüüdisajal võib näiteks täheldada selget vastavust ühelt poolt majanduse kasvu- ja konkurentsivõime ning teiselt poolt nende õpilaste arvu kasvutempo vahel, kes jätkavad õpinguid — teadmiste omandamist — pärast 15. eluaastat. Nagu märgib P. Drucker, on see näitaja kõige parem Nõukogude Liidul, Ameerika Ühendriikidel ja Jaapanil.

Võib olla kindel, et veelgi selgemalt annab informatsiooni tootmine ilme tuleviku tsivilisatsioonile. Nagu ennustatakse, kujunevad tuleviku hiigellinnad «megapolised» vaimse töö tegijate ümber, mitte tööstusettevõtete ümber, nagu see oli möödunud sajandi lõpul ja käesoleva sajandi alguses. Megapoliste eripäraks pole vabrikukorstnad, vaid teaduslinnakud. Informatsioon kujutab endast nende peamist toodangut ja ka põhilist tarbimisobjekti.

Nõukogude Liidu territooriumil elavate inimeste arv on kahekordistunud umbes 70 aastaga: 1897. aastal oli neid 124 miljonit, 1968. aastal 237 miljonit (1978. aastal 260 miljonit). Teadusega tegelevate inimeste arv kahekordistub aga iga 10 aastaga või veelgi rutem. 1950. aastal oli Nõukogude Liidus 162 500 teadustöötajat (koos kõrgkoolide õppejõududega), 1960. aastal 354 200, 1965. aastal juba 664 600, 1975. aastal 1 223 400.[2] See näitab, missuguses tempos kasvab ideede ja teadmiste tootjate protsent ühiskonnas.

Niisiis on tarvis toota üha rohkem ja üha kiiremini uut informatsiooni, et progressi ülal hoida. See on ühiskonna arengu vältimatu seaduspärasus.

Ent siit algavadki probleemid, mis viimastel aastakümnetel end üha valusamini tunda annavad. Neist rääkides kasutatakse enamasti termineid *infoplahvatus, infouputus, infokriis.* Lühidalt on probleem selles, et liiga palju on informatsiooni, samal ajal kui selle tarbijatel on liiga vähe aega ja võimalusi selleks, et mingi küsimuse suhtes olulist informatsiooni välja selgitada, kätte saada ja läbi töötada.

Viimase poole sajandi jooksul on teadus- ja tehnikapublikatsioonide hulk kahekordistunud keskmiselt iga 20 aastaga, lähematel aastakümnetel aga kahekordistub see ilmselt iga 10 aasta järel või veelgi kiiremini.

Publikatsioonide ilmumise tohutu tempo tõttu ei suuda juba praegu ükski teadlane üles leida ja läbi lugeda kõike, mis tema erialal il-

[1] Drucker P. F. The Age of Discontinuity. New York, 1968, p. 263.

[2] Народное образование, наука и культура в СССР. М., 1977.

mub. Kui uskuda J. Šeininit, siis antakse maailmas iga päev välja 100 trükipoognat kirjandust iga teadus- või tehnikaspetsialisti jaoks. See on kümme korda rohkem kui spetsialist oleks suuteline läbi lugema, isegi kui ta kulutaks kogu aja ilmunud tööde lugemisele.[3]

Ent informatsioon ei ole probleemiks ainult teaduses ja tehnikas, vaid ka majanduses, eelkõige majanduse juhtimises ja organiseerimises. Mingi otsuse vastuvõtmine kuitahes üldises või konkreetses küsimuses eeldab, et otsuse tegija käsutuses oleks võimalikult täielik informatsioon olukorra kohta, mida otsus puudutab. Mida puudulikum on informatsioon, seda suurema tõenäosusega on otsus ühes või teises suhtes defektne ning võib viia tagajärgedeni, mida keegi pole ette näinud ega soovinud.

Tootmis- või ärijuht peab konkreetsete otsuste tegemiseks pidevalt läbi vaatama ja arvesse võtma andmete tohutu hulga, mis ei kasva aeglasemas tempos kui teadus- või tehnikainformatsiooni hulk. M. Meltzeri andmetel saabus 1971. aastal Ameerika Ühendriikide äriasutustesse 15 triljonit (15 000 000 000 000) dokumenti — kirju, aruandeid, korraldusi, tellimusi jne. Iga tööpäeva igas minutis lisandus sellele paberiuputusele miljon lehekülge uut materjali.[4]

Informatsiooni h u l g a tohutu kasv, mida teadlane, tootmisjuht või ametnik peab päevast päeva läbi töötama, ei ole aga kaugeltki ainus probleem, mida peetakse silmas infokriisist rääkides. Infohulga plahvatuslik kasv on kaudselt tekitanud mitmeid muid raskusi, mis olukorra tõsidust mitmekordistavad.

Infouputuse seaduspärane kaasnähtus on informatsiooni i n f l a t s i o o n. Väga suur osa toodetavaist teadmistest osutub praktiliselt väärtusetuks — neid ei kasutata kusagil ega kunagi. Igal teadusalal on välja kujunenud omamoodi hierarhia väljaannete (eelkõige ajakirjade) osas: ühelt poolt väike hulk ajakirju, milles ilmub valdav osa selle ala olulisi töid ja mida seepärast iga selle teadusala esindaja püüab lugeda, ja teiselt poolt suur hulk ajakirju, mida loetakse alles teises või kolmandas järjekorras või ei loeta üldse.

1966. aastal leiti näiteks ühes suuremas Ukraina raamatukogus, et tellitud 2000 ajakirjast ei nõutud kolme neljandikku, s. o. umbes 1500 ajakirja, aasta jooksul ainsatki korda. 172 ajakirja läks vaja ainult ühel korral.[5]

Informatsiooni väärtuse langusel on mitmeid põhjusi. Näiteks on väga levinud komme, et autor avaldab ühe ja sama töö üsna vähe muudetud kujul mitmes eri kohas, sest tema kui infotootja tööviljakuse põhinäitajaks juhtkonna silmis on enamasti toodangu maht. Selle tulemusel informatsiooni hulga

[3] V l a d i m i r o v S., K a r e v M. Informatsioon ja meie. Tln., 1972, lk. 32—33.

[4] M e l t z e r M. The Information Imperative. New York, 1971, p. 3.

[5] V l a d i m i r o v S., K a r e v M. Informatsioon ja meie. Tln., 1972, lk. 21.

kasv ei kajasta kaugeltki meie teadmiste kvalitatiivset kasvu. Pealegi sunnib niisugune publitseerimissurve kiirustama, mille tulemusena töödesse lipsavad sisse faktivead, mis tsitaatidena kanduvad teiste autorite töödesse ja võivad informatsiooni tarbijaid väga olulisel määral desorienteerida.

Selles olukorras ei ole võimatud isegi kurioosumid — juhtumid, kus tõelise informatsiooni pähe avaldatakse ja võetakse tõena lausa mõttetusi. M. Meltzer kirjutab kahest Jaapani äriametnikust, kes olid firma poolt New Yorki saadetud. Nad tegelesid seal sellega, et mängisid hotellitoas mahh-jongi (hiina päritoluga mäng, mida mängitakse doominotaoliste tahvlikestega) ja koostasid selle tegevuse kohta pikki mõttetuid aruandeid, mis saadeti firmale. Selle tulemusel kaotas firma lõppkokkuvõttes kaks miljonit dollarit.

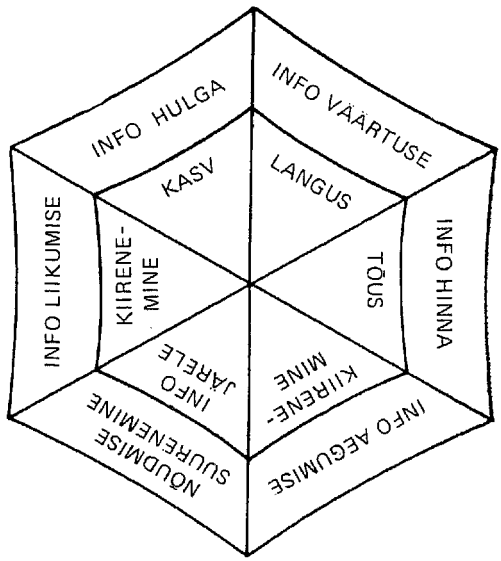
Läheb üha raskemaks leida sellisest informatsiooniprotsessist just antud juhul, antud probleemi seisukohalt vajalikke ning tõsikindlaid andmeid. Eelkõige sellest ongi tingitud paradoksaalne fakt, et samaaegselt informatsioonihulga kasvu ja tema kvaliteedi langusega kasvab kiires tempos informatsiooni h i n d. Mingi kindla koguse uue informatsiooni tootmiseks tuleb teadlasel, äri- või tootmisjuhil eelnevalt läbi töötada üha rohkem informatsiooni. Kõigepealt tuleb see informatsioon aga hankida, mis muutub aina kulukamaks. Ühes Ameerika terasekorporatsioonis lähtuti põhimõttest, et kui mingi uuring maksab vähem kui 100 000 dollarit, siis on odavam see ise ära teha kui hakata kindlaks tegema, kas samasugune uuring on kusagil juba tehtud.

Mida kiiremas tempos toodetakse uut informatsiooni, seda kiiremini ta aegub ja seda rohkem läheb maksma vaheaeg uute ideede loomise ja nende rakendamise vahel. Niisiis peab informatsioon liikuma üha kiiremini, üha operatiivsemalt. Nii paradoksaalne kui see esimesel hetkel ei tundu, nõuab selle tingimuse täitmine taas aina rohkemat informatsiooni. «Vaheaeg informatsiooni tootmise ja tarbimise vahel läheb üha rohkem maksma. Informatsioon peab voolama kiiremini kui kunagi varem. Kiired muutused, suurendades uute ja ootamatute probleemide hulka, suurendavad teiselt poolt vajamineva informatsiooni hulka. Selleks et toime tulla täiesti uue ülesande lahendamisega, on vaja hoopis rohkem informatsiooni kui selleks, et toime tulla probleemiga, mida oleme tosin või sada korda varem lahendanud. Just see nõudmine üha rohkema ja kiirema informatsiooni järele on see, mis meie ajal on hakanud õõnestama suuri vertikaalseid hierarhiaid, mis on nii iseloomulikud bürokraatiale.»[6]

Eespool puudutasime ainult põhimõttelisi probleeme, mis on esile kerkinud seoses infoplahvatusega. Kokkuvõtlikult võib olukorda kujutada «ämblikuvõrguna», mis on ära toodud joonisel 1.

Kuid mis siis õieti toimub? Mil-

[6] T o f f l e r A. Future Shock. New York, 1970, p. 121.



Joon. 1

lest kirjeldatud olukord on tingitud? Miks on inimestel vaja toota tohutul hulgal informatsiooni, kui nad ise seda tarbida ei suuda? Või kui niisugust informatsioonikogust on siiski tarvis, siis millega seletada, et tarbida ei suudeta? Lõpuks — milleni viib asjade niisugune areng, milles seisneb väljapääs olukorrast?

Kirjeldatud nähtusi ja protsesse võib kõiki vaadelda ühe põhiprotsessi otseste või kaudsete ilmingutena. See protsess on t e a d u s- j a t e h n i k a r e v o l u t s i o o n.

Mida kõrgemale arenguastmele jõuab ühiskond, seda kiiremini kasvab selles teaduse roll teiste tegevusaladega võrreldes. Meie ajal pole teadus enam lihtsalt üks ühiskonnaliikmete tegevusalasid, ta on muutunud vahetuks tootlikuks jõuks. Nüüdisühiskond, sealhulgas materiaalsete väärtuste tootmine, ei ole enam ammu mõeldav ilma teaduseta. Kahekümnenda sajandi keskel on teaduse ja tehnika areng omandanud eriti tormilise, lausa plahvatusliku iseloomu. Seda arengufaasi nimetataksegi teadus- ja tehnikarevolutsiooniks, mis on teaduse ja tehnika arengu objektiivne seaduspärasus. Informatsiooniplahvatus, mida eespool kirjeldasime, on mainitud protsessi orgaaniline koostisosa, teadus- ja tehnikarevolutsiooni teine külg. Tähendab ju teadus eelkõige ideede ja

teooriate loomist, tehnika aga kujutab endast selle informatsiooni üht teostust praktikas.

Üldu üksnes näitab kätte laiema konteksti, milles tuleb otsida vastust eeltoodud küsimustele, kuid ei anna iseenesest neile mingit vastust. Pigem vastupidi: ettevaatlikult või pessimistlikult meelestatud inimesel on nüüd alust neid küsimusi laiendada kogu teadusele ja tehnikale. Kas meil on täielik ettekujutus sellest, kuhu lõpuks viib teaduse ja tehnika tormiline areng, sealhulgas informatsiooni mahu pidev kasv, vajadus teda üha kiiremini töödelda ja sellest omakorda tulenev vajadus teda veelgi kiiremini juurde toota? Oleme harjunud suhtuma teadusesse kui kõikvõimsasse jõusse, mida on nimetatud «XX sajandi suurimaks lootuseks». Kuid kas me suudame ette näha selle jõu pidurdamatu kasvu kõiki sotsiaalseid ja psühholoogilisi tagajärgi?

Ilmeka pildi teaduse üha kiireneva arengutempoga seoses tekkivaist kahtlustest annab ameerika aatomifüüsiku R. Lappi järgmine kirjeldus: «Mitte keegi, ka mitte kõige väljapaistvamad nüüdisaja teadlased ei tea kindlalt, kuhu meid viib teadus. Me nagu asuksime rongis, mis aina kiirust juurde võttes kihutab mööda raudteed, kus rohked haruteed viivad teadmata suundades. Masinistikabiinis ei viibi ainsatki teadlast, juhtimispuldi juures seisavad aga võib-olla deemonid ise. Enamik reisijaid istub viimases vagunis ja vaatab tahapoole. Mõned neist, haaratuna kahtlustest, et on istunud põrgusse viivale ekspressile, tahavad maha hüpata, kuni pole veel hilja. Ilmneb aga, et niisugust võimalust neil enam pole.»[7]

Paljud teadlased ja ühiskonnategelased on hakanud rääkima «teaduse emantsipatsioonist», pealetungist teistele inimese vaimse tegevuse aladele, kuivõrd teaduslik, «objektiivne» tunnetus asetatakse kõrgemale kõigist muudest inimtunnetuse vormidest. Nii on näiteks šveitsi filosoof A. Mercier avaldanud arvamust, et teadus selle tänapäevasel «tehniseeritud» kujul kujutab endast ohtu inimkultuurile. «Tarkus redutseeritakse teaduseks. Teaduslik kultuur kitsas mõttes triumfeerib nn. humanitaarse kultuuri üle. Objektiivsus on asetatud mõõtmatult kõrgemale subjektiivsusest, kaemuslikkusest. Eriti on kannatada saanud kontemplatsioon, mõtisklev süvenemine, sisevaatlus. Teadus on läinud agressiivseks ning püüab vaimuelus haarata juhtpositsiooni. Ta võimutseb, muutub monopoliks, kergekäeliselt kõike muud ignoreerides. Just seepärast, et inimestel jääb puudu mõtisklevast süvenemisest, kaemusest ja tõelisest kunstist, pöörduvad nad surrogaatide poole kuni narkootikumide ja pornograafiani välja. Kokku võttes tallatakse tarkus jalge alla.»[8]

Midagi samasugust on väitnud juba O. Spengler käesoleva sa-

---

[7] Lapp R. The New Priesthood. New York, 1965, p. 29.

[8] Mercier A. Teadus on muutunud agressiivseks. — Sirp ja Vasar, 1975, 17. jaan.

jandi algul oma tuntud raamatus «Õhtumaa allakäik»[9], kirjeldades kaasaegset ajajärku kui Lääne kultuuri lõppu, konkreetsemalt kui selle üleminekut tsivilisatsiooniks. Vastandina kultuurile, mille tuuma moodustab kunstide ja filosoofia arenemine, kujutab tsivilisatsioon Spengleri järgi endast sellist ajastut ühiskonna arengus, mis tugineb üksnes teadusele ja tehnikale, mille arendamisel juhindutakse puhtutilitaarsetest eesmärkidest. Materiaalsed väärtused domineerivad täielikult eetiliste ja esteetiliste üle. Inimkond ei arene enam sügavuti, vaid laiuti. Tema sümboliks on Faust, kelle hinge on lõplikult vallutanud mefistolik alge. Inimene osutub kõige selle vangiks, mis ta ise on loonud; ta muutub olendiks, kes sedamööda, kuidas kasvavad tema teadmised, üha vähem käsutab ise oma saatust.

Ei ole siiski eriti raske tõestada, et viimati kirjeldatud seisukohad oma teadusevaenuliku hoiakuga on selgelt ühekülgsed (nagu muidugi on ühekülgsed ka need seisukohad, mis seostavad inimkonna progressi *ainult* teaduse ja tehnika arenguga).

Peab ausalt tunnistama, et me tõepoolest ei suuda praegu kuigivõrd täpselt ennustada, kuhu teaduse ja tehnika areng lõpuks välja jõuab, ning oleme juba praegu tunda saanud mitmeid selle arengu negatiivseid kaasnähtusi. Kuid see pole kaugeltki küllaldane põhjus, et teadus tervikuna hukka mõista ning kuulutada ta inimkultuurile ohtlikuks nähtuseks. On mõttetu esitada vastukäivaid nõudmisi teaduse ja tehnika progressile: nõuda, et see looks jõude, mis suudaksid purustada mägesid ja toita miljardeid inimesi, ja samal ajal tahta, et need jõud ei oleks ohtlikud kärbselegi. Kui inimkond vajab ja tahab käsutada niisuguseid jõude — ja neid ta kahtlemata vajab —, siis peab ta oskama neutraliseerida ka nendega kaasnevad ohud.

Peamine on aga, et miski peale teaduse pole suuteline teenima üha suureneva inimkonna üha kasvavaid vajadusi — ja midagi pole selleks pakkuda ka teaduse kriitikutel.

Aktiivne suhtumine ümbritsevasse maailma, selle eesmärgipärane uurimine ja ümberkorraldamine — suhtumine, mis kulmineerub teaduses — on olnud ja on kogu meie nn. euroopaliku tsivilisatsiooni alus. Ja selles on ka kõigi tema saavutuste seletus, sealhulgas seletus asjaolule, et tänapäeval on «euroopalik dünaamiline tsivilisatsioon» kogu maailmale arengu eeskujuks ja näidiseks. Rahvad, kes võiksid uhkustada palju vanema ja palju keerulisema kultuuriga, on sunnitud importima moodsat tehnoloogiat noorema ja lihtsama kultuuriga maadest.

Siin on ilmselt eriti sobiv Stanislav Lemi järgmine arutlus. Lem ütleb, et kõige selle kurja juures, mis teadus meile on teinud, ei saa siiski unustada, et just teadus on päästnud suurema osa inimkonda

[9] Spengler O. Der Untergang des Abendlandes. I—II. München, 1922—1924.

näljaeksistentsist. Samal ajal Idamaade müstilised religioossed doktriinid (näiteks budism), mida võib vaadelda näitena väärtussüsteemidest, mis sisendavad põlgust ratsionaalse mõtlemisviisi, tehnika ja katseliste uuringute vastu ning seavad kõigest kõrgemale müstilise kaemuse ja kontemplatsiooni, on hoidnud sealseid vana tsivilisatsiooniga maid pikki sajandeid praktiliselt surnud seisus. «Ainult tänapäeva tööstuslik ja bioloogiline tehnoloogia suudab toime tulla massitsivilisatsiooni probleemidega, kuna kõigi aasialike doktriinide vundamendiks on... ükskõiksus — samavõrd ülev kui katastroofiline oma tagajärgede poolest —, ükskõiksus massiprobleemide vastu, pidevalt kasvava inimkollektiivi probleemide vastu.»[10]

On vaid tarvis endale selgelt aru anda, mille poole me püüdleme, jätkab Lem teisal. Kui meie püüdeks ei ole muuta maailma, vaid iseendid, ja sedagi üksnes eesmärgiga võimalikult paremini kohaneda tegelikkusega isikliku eksistentsi lühikeseks perioodiks, siis polegi budism kõige halvem valik. Kui aga seame kõigest kõrgemale teesi «enim hüvesid suurimale arvule inimestele», siis ei saa mingid religiooni, «müstilise kaemuse» ja muu sellise eetilised ja esteetilised väärtused varjata meie eest fakti, et maailma täiustamise vahenditena on nad kõlbmatud. Nad on sama arhailised nagu loosung «tagasi looduse juurde».

Oma ainevalla juurde tagasi pöördudes võime neist üldistest arutlustest teha ilmse järelduse, et lahenduse teadus- ja tehnikarevolutsiooni poolt sünnitatud probleemidele, nende seas ka infokriisist tulenevatele probleemidele peab leidma ja saab leida üksnes teaduse enda raames. Teaduse arengu kunstlik peatamine või aeglustamine on mõeldamatud. Teaduse plahvatuslik areng nüüdisajal ei ole inimkonna õnnetus, vaid on talle eluliselt vajalik; sellele rajanevad tema kõige suuremad tulevikulootused.

Asume nüüd taas infokriisi põhiprobleemi juurde: ühelt poolt on vaja üha rohkem informatsiooni ja on vaja seda üha kiiremini rakendada, et tegelda teadusliku ja tehnilise loominguga, tulla toime ühiskonna ja eriti majanduse juhtimisega; teiselt poolt aga kasvab loodav kui ka vajaminev informatsioonihulk niivõrd, et inimesed ei ole enam võimelised seda vajaliku kiirusega töötlema. Niisuguses formulatsioonis tuleb selgelt esile ka kujunenud kriisiolukorra peasüüdlane, nõrk lüli informatsiooni ringluses, ja see on inimene ise.

Tänu tootmise mehhaniseerimisele ja eriti automatiseerimisele on inimene saavutanud, et protsesside ulatus, toodetavad aine- ja energiakogused ning nende rakendamine ei sõltu kuigivõrd tema, s. t. inimese füüsilistest võimetest. Kuid protsesside tohutud mastaabid ja keerustumine nõuavad üha täiuslikumate tehnoloogiate väljatöötamist, nende üha operatiivsemat juhtimist ja kooskõlastamist.

[10] Лем С. Сумма технологии. М., 1968, с. 170.

Automatiseerimise läbi on inimene otseselt tehnoloogilistest protsessidest väga sageli eemaldunud, ta ei ole enam nende obligatoorne lüli. Kuid informatsiooni tootmise ja töötlemise protsessides, mis moodustavad tootmise juhtimise ja plaanimise, nagu ka arengu aluseks oleva teadusliku ja tehnilise loomingu sisu, pole inimene seni ainult üks lüli teiste hulgas, vaid peamine, vältimatu lüli. Kogu süsteemis ringlev informatsioonikogus peab käima läbi inimese. Inimene seab tootmise eesmärgid, hindab ja võtab vastu otsused ühe või teise tegutsemisvariandi sobivuse kohta nende eesmärkide saavutamiseks, inimene üksi tegeleb teadusliku ja tehnilise loominguga. Nende ülesannete täitmiseks peab aga inimene — riigi- või tootmisjuht, teadlane, insener — olema kursis kogu vastavat ala puudutava informatsiooniga. Ta peab selle informatsiooni kõigepealt hankima — kuid seegi ei ole kaugeltki lihtne, nagu eespool nägime. Tuleb informatsioon läbi töötada, anda talle vajalik vorm ning selle põhjal vastu võtta vajalikud otsused, välja töötada nõutavad projektid, plaanid või teoreetilised kontseptsioonid. Teisiti öeldes — tuleb luua uus informatsioon.

Eespool kujutatud olukord sunnib meid tunnistama, et inimesed lihtsalt ei tule toime kirjeldatud ülesannete täitmisega. Ainus võimalik väljapääs on asuda ka infotöötlusprotsesside radikaalsele automatiseerimisele. See on juba teine tasand: ei automatiseerita vahetuid tehnoloogilisi protsesse, vaid nende juhtimiseks vajalikku intellektuaalset infotöötlust. Kõik protsessid, mida eespool kirjeldasime, ei nõua inimese osavõttu mitte tema füüsiliste, vaid intellektuaalsete omaduste pärast. Inimene on praegu nende protsesside vältimatu lüli seepärast, et seni omab ainult inimene nõutava tasemega intellekti. Seega pole jutt lõppkokkuvõttes vähemast ega rohkemast kui inimese intellektuaalse tegevuse automatiseerimisest.

«Tänaste «teise tasandi» automatiseerimisotsingute mõte pole niivõrd inimese füüsiliste võimete suurendamises, tema käte funktsioonide asendamises, kuivõrd selles, et automaatidele üle kanda inimese informatsiooniline tegevus — tema olukorra tajumise oskus, märkide mõistmine (seejuures selliste märkide mõistmine, mis on spetsiifilised inimesele, nagu kõneldav keel, käekiri, joonistus), tema oskus arutleda (ja seejuures õigesti), kõrvutada, hinnata, püstitada eesmärke, võtta vastu otsuseid ja leida teid nende elluviimiseks.»[11] Ainult inimese eelkirjeldatud funktsioonide automatiseerimine võib olla infokriisist tekitatud probleemide lahenduseks. Ainult sedamööda, kuidas inimesel õnnestub automatiseeritud süsteemidele üle anda oma intellektuaalse tegevussfääri üksikud lülid kuni kõige keerulisemate, kõrgemate mõtlemisprotsessideni välja, võib inimene ise vabaneda infotulva survest.

[11] Полетаев В. Предисловие. — В кн.: Человеческие способности машин. М., 1970, с. 6.

Infotulvast on palju kirjutatud, nagu ka infotöötluse automatiseerimisest. Kuid tuleb silmas pidada, et inimkond on alles nende probleemide lahendamise algusjärgus. Töö füüsilise ja intellektuaalse komponendi üleandmine masinatele on seni kulgenud äärmiselt ebaproportsionaalselt, esimest komponenti eelistades. Tegelikkuses, tööprotsessis endas, peab nende komponentide vahel aga valitsema kindel suhe. Ju siit algabki kõik see, mida tavatsetakse nimetada «tsivilisatsiooni koormaks»: inimesed elavad üha tugevama intellektuaalse ja emotsionaalse pinge all, nad aina tormavad, aga ometi ei jõua teha kõike, mida oleks vaja teha. Surub peale «füüsiline komponent», mille oleme nii võimsaks paisutanud ja mis nüüd nõuab meilt endilt vajalikus proportsioonis «intellektuaalset komponenti». Eks ole kirjeldatud olukord see, mis annab tõe kõla nende sõnadele, kes kuulutavad, et inimesed on muutunud kõige selle orjaks, mis nad on loonud, selle asemel, et saada selle isandaiks. Võib arvata, et alles siis, kui masinad suudavad võimendada inimese intellektuaalset suutlikkust samal määral kui tema füüsilisi võimeid — niisiis kümneid tuhandeid kordi —, võib loota tõelise tasakaalu taastumisele nii ühiskonnas tervikuna kui ka iga üksiku liikme elus.

I.2. Infokriis, arvutid ja keel — keeleline sissejuhatus

Eelmises jaotises sai selgeks, et kui mitte ainuke, siis peamine infokriisist jagusaamise tee on intellektuaalse infotöötluse automatiseerimine, selle õpetamine arvutile. Käesolevas jaotises püüame selgitada kaht asja. Esiteks, miks infotöötluse automatiseerimine selles mõttes, nagu me temast eespool rääkisime — kui infoplahvatuse inimese kontrollile allutamise vahend — peab haarama ühe olulisema alana loomuliku keele tekstide töötlemist, nende moodustamise ja mõistmise automatiseerimist. Ja teiseks, mis on selles ülesandes nii keerulist, et teda pole senini suudetud lahendada, olgugi et ülesanded on saanud selgeks juba ammu ning on enam kui pakilised.

Otsene vastus esimesele küsimusele tuleneb vahetult kõigest sellest, mida rääkisime seoses infouputuse ja infokriisiga. Valdavas osas on see informatsioon sõnastatud mingis loomulikus keeles, on t e k s t i n f o r m a t s i o o n. Teadusliku ja ka tehnilise informatsiooni puhul on see ilmne. Raamatud, artiklid, aparaatide või tehnoloogiate kirjeldused — kõik need on tekstid. Ent ka majanduse ja ühiskonna juhtimisel ei moodusta sõnaline informatsioon sugugi väikest osa. Piisab ühest näitest: seadused ja muud õigusaktid, mida igal vähegi arenenud maal kehtib kümneid tuhandeid, on kõige tava-

lisemad tekstid. Infouputus, millest rääkisime, on sellelt seisukohalt vaadatuna tekstiuputus — teadlase jaoks erialaprobleeme käsitlevate artiklite ja raamatute uputus, tootmisjuhi või administratiivtöötaja jaoks dokumendiuputus ehk lihtsalt «paberiuputus». Et arvuti suudaks aidata inimesel infouputusest jagu saada, peab ta olema suuteline tekstidega vajalikul viisil opereerima. Probleemiks ei ole seejuures mitte ainult vajalike tekstide läbilugemine. Enne on tarvis otsustada, missugustes tekstides üldse sisaldub vajalikku informatsiooni ja missugustes ei sisaldu, on vaja teada saada, kus antud küsimuse kohta on varem midagi kirjutatud või ei ole, ja kui on, siis kas seda tasub lugeda või mitte. Arvuti peab aitama informatsiooni otsida, peab suutma refereerida artikleid ja raamatuid, koostada bibliograafilisi ja temaatilisi ülevaateid ja palju muudki (mida nimelt, selle juurde pöördume tagasi peatüki lõpus).

Kuid lisaks sellele silmanähtavale põhjusele, miks on tarvis õpetada arvutit keelest aru saama, on veel teine, sügavam ja üldisem põhjus, õigemini põhjuste kompleks, millel tasub peatuda pisut lähemalt.

Inimese intellektuaalse tegevuse üleandmist automatiseeritud süsteemidele, millest kõnelesime eelmises jaotises, ei tule mõista nii, et arvutid teevad tulevikus ära inimeste kogu mõistusliku töö teaduses, tootmise ja ühiskonna juhtimisel ja kõikvõimalike muude ülesannete lahendamisel, ning et inimene neist protsessidest täielikult välja lülitatakse. Vaevalt niisugune olukord kunagi saabub ja selle poole me ei püüagi. Ehkki on vajalikud ka iseseisvalt tegutseda võivad intelligentsed süsteemid, näiteks robotid, ei ole nende loomine arukate süsteemide väljatöötamise peaeesmärk. Eelkõige on vaja, et arvuti lülituks teaduslikku ja tehnilisse uurimistöösse, majanduse ja ühiskonna juhtimisse ning plaanimisse inimese partnerina, assistendina. Ta ei asenda täielikult inimest, aga ta ei pea ka olema täpselt formuleeritud ülesannete rutiinne lahendaja. Arvutitest on kõige rohkem kasu, kui nad abistavad inimesi kõigil etappidel: probleemi esialgsel püstitamisel, selle täpsel formuleerimisel, optimaalse lahenduse väljatöötamisel, vajalike andmete kiirel ülesotsimisel ja töötlemisel. On arusaadav, et inimese ja arvuti niisugune sümbioos eeldab pidevat ja operatiivset suhtlemist, dialoogi partnerite vahel. See omakorda eeldab tõhusa suhtlemisvahendi olemasolu.

Mida enam selle probleemiga on tegeldud, seda kindlamalt jõutakse veendumusele, et ideaalne oleks arvutiga suhtlemiseks kasutada loomulikku keelt. Inimese mõtlemine on keelega otsekui kokku kasvanud. Eriti käib see loova mõtlemise, uute ideede otsingule suunatud mõtlemisprotsesside kohta, täpsemini nende sisu väljendamise kohta. Sundida inimest väljendama oma loomingulisi ideid, eriti aga nende kujunemist, arutlusi ja vahetulemusi mingis täpses formaalkeeles tähendaks sundida teda suu-

rel määral loobuma nende protsesside spetsiifikast. Ideid, mis on veel lõplikult kristalliseerumata, kuid mille tuum on inimesele juba sedavõrd selge, et neid saab teistega arutada, võimaldab väljendada ainult loomulik keel.

Juba rohkem kui viisteist aastat tagasi formuleeris selle mõtte täie selgusega üks nõukogude tuntumaid programmeerimisspetsialiste A. Jeršov: «... kui tahame, et arvuti kujuneks inimese tõeliseks partneriks tema vaimses tegevuses, siis on loomulik, et ta peaks koostöös osalema süsteemina, millele on antud niivõrd «inimesestatud» vorm kui võimalik, s. o. süsteemina, mis on suuteline olema pidevas ja vahetus kontaktis inimesega, mis omab rikast ja paindlikku struktuuri, võimelist kohanema mis tahes konkreetse situatsiooniga, milles inimene töötab, ja mis võib vastu võtta juhtnööre kujul, nagu need inimesele endale pähe tulevad, nimelt ehtsa inimliku vestluse vormis.»[12]

Paljugi sellest, millest käesolevas raamatus juttu tuleb, muutub arusaadavamaks, kui vaatame pisut lähemalt, kas soovi taga suhelda arvutiga loomulikus keeles on ainult inimlik mugavus või on tegemist hoopis sügavamal peituva vajadusega. Vaatleme küsimust kahest aspektist.

Praegu taandub arvutiga suhtlemine enamasti sellele, et inimene — reeglina mitte see, kes vahetult on huvitatud vastava ülesande lahendamisest, vaid eriettevalmistuse saanud matemaatik-programmist — koostab programmi, kuna arvuti, teostades täpselt programmis ettenähtud operatsioonid, lahendab ülesande ja väljastab tulemused. Programm on erilises programmeerimiskeeles kirjapandud käsujärjend, mis on koostatud nii, et nende käskude täitmine (vastavate operatsioonide teostamine) arvuti poolt tagab ülesande õige lahendamise. Eri programmeerimiskeelte arv ulatub sadadesse ja nende abil on võimalik lahendada väga erinevat tüüpi ülesandeid. Kuid peaaegu kõik senised programmeerimiskeeled on sarnased selle poolest, et nad on mõeldud vastavate ülesannete lahendusalgoritmide formaalselt täpseks esitamiseks. Ülesande formuleerimine, selle põhimõttelise lahenduskäigu valik ja vastava algoritmi koostamine — kõik see jääb tavaliselt inimese hooleks ja peab olema tehtud enne programmi kirjutamist; alles programmi valmimisel saab pöörduda arvuti poole.

Kuid on selge, et arvutite rakendamiseks eespool kirjeldatud viisil ei ole niisugune suhtlusviis kaugeltki rahuldav. Et automatiseeritud süsteem töötaks inimese partnerina ülesannete lahendamise kõigil etappidel, on vaja, et sellele saaks esitada mitte ainult (ja mitte niivõrd) täpseid eeskirju eelnevalt formaliseeritud andmete töötlemiseks. On vaja, et saaks esitada ka ülesandeid nende esialgses mitteranges sõnastuses, enam-vähem nii, nagu seda teeb inimene teisele

[12] Yershov A. P. One view of man-machine interaction. — Journal of the ACM, 1965, vol. 12, N 3, p. 316.

inimesele vastava ülesande olemust seletades; arvuti üks funktsioone olekski sellisest kirjeldusest lähtudes välja töötada ülesande lõplik ja täpne formulatsioon ning seejärel lahendusprogramm — seda loomulikult koostöös inimesega, pärides temalt täiendavaid andmeid ning täpsustades järk-järgult ülesande sisu. Seni loodud programmeerimiskeeled on rõhuvas enamikus niisuguseks dialoogiks kui mitte päris sobimatud, siis väga kohmakad. Olukordade või probleemide sisulisi kirjeldusi, seda enam selliseid, mis pole täiesti ühemõttelised ja kõigi detailideni selged, ei ole nende abil võimalik esitada. Loomulik keel on aga just kohanenud niisuguste kirjelduste esitamiseks, sest inimeste omavaheline suhtlemine seisneb valdavalt kirjeldava informatsiooni andmises üksteisele. Ka ülesannete ja nende lahenduste üksteisele seletamisel kasutavad inimesed enamasti niisugust suhtlusviisi.

Et loomulik keel on kõige otstarbekam ja täiuslikum suhtlusvahend inimese ja automatiseeritud süsteemi vahel, selle seisukohani jõuame veelgi kindlamalt, kui vaatame probleemile pisut teisest küljest.

Elektronarvutite tungimisega üha uutele inimtegevuse aladele kaasneb kasutajate ringi laienemine ja nende arvu järsk kasv. Väljapääsu infokriisist suudavad niisugused süsteemid pakkuda vaid sel juhul, kui nad on kättesaadavad igaühele, keda infotöötlusega seotud probleemid puudutavad — igale teadlasele, insenerile, tootmisjuhile.

Ent kui suhtlemiseks arvutiga kasutada tavalisi programmeerimiskeeli, eeldab see süsteemi kasutajalt väga suurt eeltööd oma konkreetse probleemi arvutile esitamiseks — rääkimata sellest, et eeltööga toimetulemiseks on vaja eriettevalmistust programmeerimises. Seetõttu on veel praegugi üsna levinud suhtlemisviis arvutiga selline, kus arvuti kasutaja ei esita ülesannet otse arvutile, vaid programmistile, tehes selle talle selgeks «inimlike» vahendite ja meetoditega, ning alles too koostab vajalikud programmid, tehes sel teel ülesande mõistetavaks arvutile.

Kohmakuse ja aegluse tõttu pole niisugusel suhtlemismoodusel tulevikus suurt perspektiivi. Ühelt poolt ei ole erialaspetsialistil kaugeltki lihtne kirjeldada oma probleeme niivõrd täpsel ja formaalsel kujul, et programmist neist ühemõtteliselt aru saaks. Teiselt poolt ei suuda isegi kogenud programmist alati arvesse võtta eriala spetsiifikat (eriti kui tal tuleb tegelda mitme eriala ülesannetega), mistõttu programmid ei ole sageli kaugeltki optimaalsed.

Pole mingit kahtlust, et just siin peitub praegu üks infotöötluse automatiseerimise nõrku lülisid. Tehniliste võimaluste poolest suudaksid elektronarvutid juba praegu enamat, kui nad teevad.

Püüd leida väljapääsu kirjeldatud olukorrast ning saavutada paremat vastavust ülesannete sisu ja programmide vahel on tinginud uut tüüpi spetsialistide kujunemise, kes on vastava ala head tundjad, ent omavad ka vajalikke teadmisi pro-

grammeerimisest, nii et suudavad ise oma probleemid arvutile esitada. Ometi on programmistioskuste omandamine eriaspetsialistidele praeguste programmeerimiskeelte ja programmeerimistehnika suhtelise keerukuse tõttu siiski koormav. Pealegi võib kahe eriala sümbioos ühes inimeses anda häid tulemusi vaid juhul, kui oma ainevalla hea tundja on samal ajal ka tõepoolest hea programmist. Kui ta tunneb programmeerimist üksnes pealiskaudselt, nagu see on enamasti paratamatu, ei ole tulemused palju paremad kui programmisti vahelülina kasutamisel.

Siit jõuame taas veendumusele, et suure ja olulise ülesannete ringi puhul ei ole loomulik keel mitte ainult kõige mugavam, vaid ka kõige otstarbekam suhtlemisvahend inimese ja elektronarvuti vahel. Ta ei nõua inimeselt eri ettevalmistust ning lubab mõtteid formuleerida, kuid samal ajal määratleda ka mõisteid kui tahes täpselt ning rakendada ükskõik kui keerulisi formalisatsioone.[13]

Nüüd aga on paras aeg pöörduda vaadeldava probleemi teise tahu poole ja küsida, *mida* on keeles nii keerulist, et selle töötlemine arvutite abil on osutunud niivõrd raskeks? On ju selle probleemi lahendamisele asumisest möödunud enam-vähem kolmandik sajandit. Ometi pidi näiteks Anthony Oettinger, keele automaattöötluse suuremaid spetsialiste ning Ameerika Arvutustehnika Assotsiatsiooni omaaegne president, veel mõni aeg tagasi tunnistama: «Mida rohkem aega ma kulutan keele uurimisele, seda suurem hämmeldus mind haarab, sest viimased kaks aastakümmet, olles seninägematult intensiivse keelealase uurimistöö tunnistajaks, on esile toonud peamiselt vaid selle, missuguse hämmastava kergusega inimeste keeleline kommunikatsioon saadab korda oma igapäevaseid imesid. Olles niivõrd lihtne, et iga laps omandab vastava oskuse peaaegu ebateadlikult, on see protsess samal ajal ometi niivõrd keeruline, et kõik meie püüded teda mõista näivad andvat meile ainult fragmentaarseid ja pealiskaudseid teadmisi.»[14]

Mis on siis see, mis segab elektronarvuteid samasuguse kerguse ja enesestmõistetavusega aru saamast loomuliku keele tekstidest ja väljendamast samas vormis lahendus- ja arvutuskäikude tulemusi?

Need tegurid on loomuliku keele mõned üldomadused: väljendite ebamäärasus, mitmemõttelisus,

[13] Mitmeis programmeerimissüsteemides kasutatakse ulatuslikult nii loomuliku keele sõnavara ja konstruktsioone kui ka muid harjumuspäraseid probleemiesitusviise, näiteks matemaatiliste avaldistena; esiletõstu väärivad Eesti NSV-s tehnikadoktor Enn Tõugu juhtimisel arendatavad tööd. Sellele vaatamata peab autori poolt õeldu põhiliselt paika, sest arvutitega suhtlemises on mahult esikohal universaalsed programmeerimiskeeled, mis loomulikust keelest on üsna kaugel. Siiski leidub nendegi hulgas mis tahes eriala inimesele üsna lühikese ajaga õpitavaid keeli, näiteks BASIC. (*Toim.*)

[14] Oettinger A. Language and Information. — American Documentation, 1968, N 3, p. 295.

asjaolu, et (vähemalt näiliselt) puuduvad ranged reeglid, millest juhinduda väljendite moodustamisel ja analüüsimisel. Nende omaduste põhjusi lähemalt vaadeldes ja süstematiseerides märkame rida loomuliku keele konkreetseid jooni, millest igaüks lisab keele üldisele ebamäärasusele mingi aspekti.

Esiteks: mitmetähenduslikkus. Formaalkeelte puhul, millega elektronarvutid on harjunud opereerima, peab iga väljend võimaldama ainult üht tõlgendust. Loomulik keel on aga mitmetähenduslikkusest sõna otseses mõttes läbi imbunud.

Mitmetähenduslikkus hakkab peale juba üksiksõnadest. Vaevalt leidub eriti palju sõnu, millel oleks üks ja ainult üks kindel tähendus. Sõnaga *tulema* seostame esmajoones ruumilise liikumise, s. o. tähenduse, mis on sel sõnal näiteks lauses *Poiss tuli koolist koju.* Kuid *tulema* võib esineda ka teistes tähendustes, näiteks lausetes *Sellest poisist tuleb hea töömees, Talle tuli mõte teha kõike vastupidises järjekorras, Töö tuleb lõpule viia.*

Vähemalt samavõrra mitmetähenduslikud kui sõnad võivad olla ka grammatilised vormid ja neil põhinevad konstruktsioonid. Näiteks nimisõna omastav kääne ja vastav kahe nimisõna ühend võivad väljendada väga erinevaid sisulisi seoseid nende sõnade tähenduste vahel. Ühendis *poisi õun* väljendab esimese sõna omastav kääne ilmselt omamis- ehk kuuluvusvahekorda: poiss on õuna omanik; ühendis *lapse nutt* väljendab see vorm aga subjektivahekorda: laps on see, kes nutab; ühendis *lapse karistamine* on laps aga ilmselt see, keda karistatakse, niisiis väljendab omastav kääne hoopis objektivahekorda. Ja nii edasi. Selliseid kahe nimisõna ühendeid põhjalikult uurinud keeleteadlased on leidnud, et kõnealuse konstruktsiooni puhul (kus omastavas käändes nimisõna laiendab teist nimisõna) võib eristada rohkem kui sadat sisulist seosetüüpi objektide või nähtuste vahel, mida vastavad sõnad tähistavad.

Sõnade ja grammatiliste konstruktsioonide mitmetähenduslikkusele lisandub tervikteadete — lausete — puhul veel üht liiki mitmetähenduslikkus: sõltuvalt kõneleja eesmärgist ning kontekstist võib üks ja sama lause anda edasi väga erinevaid sisulisi teateid. Seda liiki mitmetähenduslikkust, mida võib nimetada ka mitmemõttelisuseks, illustreerime järgmise dialoogiga. Ema noomib söögilauas Jukut: «Juku, kes see sügab endal lusikaga nina!» Juku küsib imestunult: «Aga ema, kas siis seda tehakse kahvliga?» Nagu näeme, osutub ema poolt öeldud näiliselt ühemõtteline lause tegelikult mitmemõtteliseks: Juku loeb temast välja hoopis teistsuguse teate kui see, mida lauset öeldes pidas silmas ema; seejuures on Juku poolt valitud tõlgendus puhtkeeleliselt samavõrd võimalik ja õigustatud nagu see, mille väljendamiseks ema lause ütles.

Mitte vähem kui mitmetähenduslikkus on keeles levinud vastupidine omadus: võimalus üht ja sama sisu või tähendust väljendada mit-

mel viisil, erinevate sõnade või lausetega; see omadus on sünonüümia. Peaaegu iga sõna puhul — täpsemalt, iga sõna iga üksiku tähenduse puhul — võime leida teisi sõnu, mis väljendavad sedasama või peaaegu sedasama tähendust. Öeldut võib illustreerida ülaltoodud näidetes kasutatud sõna *tulema* tähendusega. Sünonüümiks *tulema* tähendusele, mis avaldub lauses *Töö tuleb lõpule viia,* on *vaja olema* või *pidama;* tähendusele, mis avaldub lauses *Sellest poisist tuleb hea töömees,* on sünonüümiks sõnad *saama* või *kujunema* (vastavates tähendustes).

Sõnaühendite ja tervete lausete puhul on varieerimisvõimalused veelgi ulatuslikumad kui üksiksõnade puhul. A. Žolkovski ja I. Meltšuk, kes on spetsiaalselt uurinud seda keele omadust, väidavad näiteks, et sisu, mida väljendab lause *Ainult erialaste terminite suur hulk antud tekstis teeb Smithi jaoks võimatuks selle tõlkimise,* on võimalik väljendada rohkem kui miljoni sünonüümse lausega (parafraasiga). Näiteks: *Ainult erialase terminoloogia rohkus antud tekstis muudab Smithile selle tõlkimise võimatuks; Miski muu peale erialase terminoloogia rohkuse antud tekstis ei takista Smithi seda tõlkimast;* jne. Nimetatud arv ei olegi nii uskumatu, kui silmas pidada kas või seda, et igal sõnal ülaltoodud lauses on kaks-kolm, mõnel aga kuni kümme sünonüümi; juba nende kõikvõimalikud kombinatsioonid annavad üpris suure arvu. Sellele lisandub mitmesõnaliste ühendite omavaheline asendatavus *(teeb võimatuks — takistab — on põhjuseks, miks ei ole võimalik — tingib mittevõimalikkuse* jne.). Sellise kombineerimise teel saadud laused ei ole muidugi stilistiliselt samaväärsed, kuid nad on sünonüümsed selles mõttes, et väljendavad üht ja sama objektiivset fakti.

Süsteem, mille ülesandeks on loomulikku keelt automaatselt analüüsida ja mõista keeleväljendite sisu, peab seda suutma teha nii, et ta näiteks antud juhul kõigi miljoni sünonüümse lause analüüsi tulemusena saaks enam-vähem ühe ja sama sisulise semantilise struktuuri, mis on vastava fakti kirjeldus elektronarvuti jaoks.

Teiselt poolt ei pea see süsteem mitmetähenduslikkuse puhul suutma kindlaks teha mitte ainult sõnade või lausete kõik eri tähendused, vaid ka selle, missuguses tähenduses vastavat sõna või lauset antud konkreetsel juhul on kasutatud. Nii näiteks peab süsteem automaatselt kindlaks tegema verbi *tulema* eri tähendused eespool toodud lausetes ning erinevad sisulised seosed, mis kehtivad sõnade vahel nimisõnaühendites nagu *poisi õun, lapse nutt, lapse karistamine.*

Nende ülesannete lahendamine on elektronarvuti jaoks seotud suurte raskustega, mille ületamisele olid suunatud esialgsed tööd keele automaatsel analüüsimisel, eriti masintõlke raames tehtud jõupingutused.

Kuid mitmetähenduslikkus ja sünonüümia ei ole ainsad omadused, mis loomuliku keele automaattöötluse nii raskeks muudavad. Nad

on vaid kergemini ja vahetumalt tähele pandavad ning seepärast ka suhteliselt paremini tuntud.

Märgatavalt vähem uuritud, kuid hoopis sügavamal peituvate juurtega on loomuliku keele omadus, mida kõige üldisemalt võib nimetada e l l i p t i l i s u s e k s: loomuliku keele tekst on alati lünklik, selle vahetus välises vormis ei ole esitatud kogu sisu, mida tekst on mõeldud edasi andma (ja tegelikult ka annab edasi) inimesest suhtluspartnerile. Võib eristada vormilist ning sisulist ehk semantilist ellipsitilisust. Esimese näiteks olgu lause *Poiss jooksis puid tooma*, milles on mainitud kaht tegevust — jooksmist ja puude toomist. Igale eesti keele oskajale on selge, et mitte ainult jooksjana, vaid ka puude toojana on selles lauses mõeldud poissi, olgugi et vormiliselt on sõna *poiss* seotud ainult sõnaga *jooksis*. Me ei märka siin toomise subjekti mainimatajätmist, sest eesti keele vastavad reeglid võimaldavad meil selle seose rekonstrueerida. Kõrvutame eelnenu lausega *Isa käskis puid tuua,* kus me märkame kohe, et on jäetud ütlemata, k e s peab puid tooma: käesoleval juhul keele enda reeglid ei võimalda vastavat informatsiooni taastada. Kuid ka viimane lause võib olla täielikult arusaadav, kui see on öeldud vastavas kontekstis, s. o. juhul kui eelnevalt on nagunii teada, keda isa käskis. Sel juhul esindabki lause ellipsit, mida võib nimetada semantiliseks: lause vahetu, välise vormi põhjal ei ole enam võimalik puuduvat informatsiooni taastada, seda võimaldavad ainult sisulised teadmised lauses kirjeldatavast situatsioonist.

Toodud lause esindab semantilise ellipsi üht kõige lihtsamat tüüpi. Keerulisem on teine tüüp, mis on rajatud arutluskäikudele, järelduste tuletamisele vahetult teatatud materjalist. Näitena järelduste vajalikkusest, aga ka enesestmõistetavusest inimese jaoks võib kõrvutada kaht vormiliselt täiesti ühesuguse ehitusega lauset: *Poiss ei suutnud ust avada* ja *Poiss ei unustanud ust avada.* Esimesest lausest järeldame, et poiss ei avanud ust, teisest aga, et poiss avas ukse; esimese lause järel teame, et uks on kinni, teise järel teame, et uks on lahti. Tekstis, kus eeltoodud lused esinevad, ei hakata informatsiooni ukse seisundist loomulikult enam eraldi teatama ning teksti analüüsiv süsteem peab suutma selle tuletada eeltoodud lausetest, nii nagu seda teeb inimene.

Kõige keerulisemad ja raskemini kirjeldatavad on aga semantilise ellipsi tüübid, mille aluseks on inimeste omavahelise suhtlemise kõige universaalsemad omadused: teise täiskasvanud inimesega kõneldes me eeldame, et suhtluspartner on arukas teadvusega olend, evib samasugust sisemaailma ning tunnetab välismaailma samasugusena nagu meiegi, omab mälu ja loogilise arutlemise võimet ning ka kindlaid sisulisi teadmisi maailma ehituse, esemete ja nähtuste tüüpiliste omaduste ning seoste kohta. Ehk nagu praegu lühidalt tavatsetakse öelda: omab kindlat maailmamudelit. Esimesel hetkel on isegi raske kujutleda, kuivõrd olulisel

määral inimesed tavalisel suhtlemisel toetuvad neile eeldustele ja kuivõrd palju informatsiooni konkreetse teksti puhul jäetakse juurde mõelda suhtluspartneril endal.

Võtame näiteks järgmise katkendi lihtsast lastejutust: *«Kui hunti näen, söön ta ise ära!» hüüdis uljas jänesepoeg. See oli juba hoopis kentsakas. Väikesed jänkud turtsatasid, kattes suud käpakestega.* Juba neid kolme lauset omaette võetult lugedes saab võrdlemisi selge pildi olukorrast ja sündmustikust, rääkimata sellest, kui lugeda neid muinasjutu terviklikus kontekstis. Ent kui palju sellest materjalist on vahetult öeldud nende lausetega ja kui palju me järeldame sellest, mis vahetult on öeldud? Näiteks: meile on selge, et lausega *See oli juba hoopis kentsakas* väljendatakse asjaolu, mis on põhjuseks järgmise lausega esitatavale faktile — et jänkukesed turtsatasid. Kuid niisugust seost nende kahe fakti vahel pole tekstis otseselt väljendatud. Me järeldame selle sisulise seose kahe mainitud fakti vahel. Selleks kasutame mitmesuguseid üldisi teadmisi, näiteks teadmist, et naeruturtsatus võib väljendada (muu hulgas) naerja suhtumist nähtusesse kui naljakasse. Samuti järeldame eeltoodud tekstist kui endastmõistetavat fakti, et jänkukesed kuulsid seda, mida uljas jänkupoiss hüüdis, ehkki sellest üldse juttu pole. Meil pole ka raske mõista, miks jänkupoisi eeltoodud hüüd kentsakas on, mistõttu muinasjutus ei hakata sedagi seletama. Ja nii edasi. Selleks et detailselt kirja panna kõik see, mida me kõnealusest tekstist lisaks selles öeldule välja loeme, läheb ühtekokku vaja mitu korda rohkem lauseid kui need kolm, millest tekst praegu koosneb. Loomuliku keele tekstide puhul on see alati nii: vahetult nähtav (loetav) või kuuldav tekst on vaid väline vorm, hõre ja lünklik, mille peamine ülesanne on juhtida teksti vastuvõtja sügavamal peituvate struktuuride juurde.

Lõpuks tuleb kindlasti rääkida loomuliku keele iseärasusest, mis tekstide automaattöötlusel kujutab endast võib-olla suurimat raskust, kuivõrd tema aluseks olevad mehhanismid on ka puht teoreetilises mõttes veel võrdlemisi tumedad. See on võimalus kasutada keeleväljendeid ülekantud, metafoorsetes ehk kujundlikes tähendustes. Selle nähtuse üks lihtsam avaldus on fraseoloogiliste väljendite rohkus keeles. Fraseoloogiline väljend on väliselt vormilt enamasti päris harilik keelekonstruktsioon, kuid tema tähendus ei tulene selles sisalduvate sõnade otsestest tähendustest mingite kindlate reeglite järgi, vaid on neist omaette seisev tervik. Näiteks: *sai kitli peale, hakkas oma sõnu sööma, viskas vedru välja.* Automaattõlkealasesse folkloori kuulub lugu sellest, kuidas USA presidendile Dwight Eisenhowerile demonstreeriti üht tolleaegset masintõlkesüsteemi. Eisenhower andis süsteemile vene keelde tõlkida ingliskeelse lause *The spirit is strong, but the flesh is weak* (selle eestikeelseks vasteks oleks vaim on valmis, aga liha on nõder)

ning süsteem tõlkis: Спирт крепкий, но мясо испортилось.

Kuid kujundliku keelekasutuse suurim probleem ei ole siiski sellistes kindlakskujunenud väljendites. On võimalik (vähemalt põhimõtteliselt) väljendid kokku koguda ja lisada nad sõnastikule, millest süsteem saab vajaduse korral kontrollida, kas mingit väljendit võib analüüsida tavaliste keelereeglite järgi või tuleb sellega opereerida kui tervikuga.

Palju tõsisemaks ja põhimõttelisemaks probleemiks on niisugused kujundlikud väljendid, mis sünnivad inimeste suhtlemise käigus ja mida ei ole võimalik mingite loenditega ette näha. *Liivatorm: kollase kotka vihisev söös ja resivateratkuvate küüniste kõigepurustus. Liivaterad on saanud äkki tiivad, see on neid sünnitanud mägede aastamiljonite unest ärkamine.* Niiviisi kirjeldab N. Baturin liivatormi kõrbes. Ja me saame aru, mida ta selle kirjeldusega tahab edasi anda, seejuures aga mitte seetõttu, et oleme samu väljendeid kui tervikuid varem kohanud. Vastupidi, niisuguseid kujundlikke kirjeldusi loeme seda paremaks ja naudime neid seda rohkem, mida uudsemad nad on. Kuid kuidas me neid mõistame, mis reeglitele see protsess allub, seda ei oska me kirjeldada, vähemalt mitte sel tasemel, et protsessi automatiseerida. Tegelikult pole see enam keelenähtus, pigem inimintellekti kui terviku üks sügavalt juurdunud iseärasusi — tunda lõbu mõistuütlemisest ning selle mõistmisest. Õnneks — tekstide mõistmise automatiseerimise õnneks — on niisugune mõistuütlemine ainult väga piiratud ulatuses omane nendele tekstidele, mille töötluse automatiseerimine on esmase tähtsusega — teaduse, tehnika ja halduse tekstidele. Mõistuütlemine ja kujundlikkus on ilukirjanduse ning igapäevase kõne pärusala, siin aga pole automatiseerimise järele esialgu veel põletavat vajadust.

Eelnenud jutust ei tohi aru saada niiviisi, nagu kujutaksid keele mitmetähenduslikkus, sünonüümia, elliptilisus ja kujundlikkus loomuliku keele puudusi. Hoopis vastupidi: just tänu neile omadustele on loomulik keel see, mis ta on; just neist saavad alguse loomuliku keele põhiväärtused: paindlikkus, rikkus, kasutatavus mis tahes uutes situatsioonides ja mis tahes uute mõtete väljendamiseks, võimalus vajaduse korral keelde juurde tuua uute tähendustega uusi sõnu, kasutusele võtta uusi konstruktsioone jne. Seepärast ei olegi mõeldav nende omaduste poolt sünnitatud raskustest üle saada, neid lihtsalt likvideerides. Igaühel neist on keeles kindel funktsioon, mida peame mõistma ning kindlaks tegema, missuguste mehhanismide kaudu ta keelelises suhtlemises realiseerub. Keel, olles mitmetähenduslik, elliptiline, täis sünonüüme ja kujundeid, sisaldab ühtaegu vajalikke vahendeid, mis iga konkreetse kasutuse — suhtlemisfakti — puhul võimaldavad neid omadusi neutraliseerida. Seda tõendab ümberlükkamatult tõsiasi, et inimesed kasutavad keelt suhtlemiseks ja saavad seejuures üksteisest enamasti aru.

Niisiis ei saa loomuliku keele tekstide sisuliseks töötlemiseks elektronarvutite abil töötlust segavaid omadusi keelest välja rookida, vaid tuleb õpetada arvutit nende omadustega niisamuti ümber käima, nagu seda teeb inimene.

Tehtud töö käigus on jõutud olulise arusaamani, et ühtki kirjeldatud probleemi ei õnnestu lahendada teistest lahus. Näiteks katsed lahendada sünonüümia ja mitmetähenduslikkuse probleemi — mille kallal seni peamiselt vaeva on nähtud — lahus elliptilisuse ning kontekstist sõltuvuse probleemidest ei ole andnud tulemusi. Lausete mitmetähenduslikkuse lahendamiseks loodud protseduurid töötavad sageli nii, et kõigepealt selgitatakse välja kõik lause erinevad võimalikud analüüsimisviisid (s. o. tähendused) ning alles pärast seda püütakse valida antud juhul kohaseim variant, mille tähendust seejärel asutakse lähemalt analüüsima. Kuid need protseduurid osutuvad niivõrd keerulisteks ja nende abil leitavate võimalike tähenduste hulk niivõrd suureks, et suuremate tekstide analüüsimisel osutuvad nad praktiliselt kõlbmatuks. Samal ajal ei kujuta lausete mitmetähenduslikkus inimestele enamasti kuigi tõsist probleemi. Harilikult tunneme konkreetsel juhul mõeldud tähenduse ja seega ka sobiva analüüsimisviisi ära peaaegu automaatselt, ülejäänud analüüsivõimalusi aga ei märkagi. Näiteks olgu lausepaar: *Ema tõttas väikest Kaarlit tröötima. Lapse lohutamisest oli aga vähe kasu.* Selles kontekstis on sõnaühend *lapse lohutamine* meie jaoks täiesti ühemõtteline: laps on see, keda lohutatakse. Et selle sõnaühendi puhul — ja seega ka terve lause puhul, kus see ühend esineb — on võimalik veel tõlgendus, kus laps ise on lohutaja, ei tule meile pähegi. Samal ajal järgmises kontekstis näeme just seda teist tõlgitsust, kuna esimene, vastupidi, ei tule pähegi, kui sellele spetsiaalselt ei mõtle: *Väike Kaarel tõttas ema tröötima. Lapse lohutamisest oli aga vähe kasu.* Niisiis on lause *Lapse lohutamisest oli vähe kasu* iseenesest võttes mitmetähenduslik, võimaldades mitut erinevat analüüsimisviisi. Konkreetsetes kasutustes osutub ta aga praktiliselt ühemõtteliseks: me näeme kohe, otsekui automaatselt, vajalikku tõlgendust. Seda sellepärast, et me lause vormilise analüüsi puhul arvestame täies ulatuses ka meie käsutuses olevat semantilist materjali, terviklikku situatsiooni, mida eelnenud laused kirjeldavad ja mis moodustab antud lause tausta.

Nüüd, kui on üldjoontes selgunud, milleks on inimesele vaja tekstide automaattöötlust ja võimalust suhelda arvutiga omaenda loomulikus keeles, samuti ka selle teel seisvad raskused, püüame anda konkreetsema ja süstemaatilisema pildi sellest, mida praktiliselt tehakse kõnealusel alal ning millest tuleb juttu järgmistes peatükkides. Ülevaate käigus selgitame ka mõne edaspidi vaja tuleva mõiste sisu.

Tegevusala, millele käesolev raamat on pühendatud, nimetatakse kõige üldisemalt keele automaat-

töötluseks. Keelematerjali, eelkõige tekstide töötlemist elektronarvuti abil tuleb vaadelda just nimelt kui tegevusala ja mitte kui teadusharu. Vastavate automatiseeritud infotöötlussüsteemide väljatöötamine tugineb teoreetilistele kontseptsioonidele, mis algselt lähtuvad mitmest teadusharust, kuid tänapäeval on hakanud juba omandama suhteliselt iseseisva teaduse ilmet. Eesti keeles puudub sellel teadusharul kindlakskujunenud nimi. Ingliskeelses kirjanduses on juurdunud termin *computational linguistics,* mis eesti keelde otse ümberpandult oleks *arvutuslingvistika.* Selline termin on tõepoolest kasutusel, kuid vaevalt saab teda pidada õnnestunuks; tekstide ja muu keelematerjali töötlemisel on arvutustega tegemist vaid väga piiratud ulatuses. Kasutatud on veel termineid *insenerlingvistika, raalilingvistika, kompuuterlingvistika,* ka *infolingvistika.* Et terminoloogiaprobleemid meid siin otseselt ei huvita, siis lepime lihtsalt kokku kasutada käesolevas raamatus nimetusi *kompuuterlingvistika* või *arvutilingvistika.* See on teaduslik uurimisala, mis varustab keele automaattöötlust keele kirjeldus-, analüüsi- ja sünteesimeetoditega, samuti teoreetiliste üldistuste ja kontseptsioonidega.

Arvutid ja keel said esmakordselt kokku juba 1950-ndate aastate alguses, seega varsti pärast elektronarvutite laiemat kasutuselevõtmist. Kokkuviimise ajendiks oli idee kasutada arvuteid tõlkimaks tekste ühest keelest teise — m a s i n t õ l k e idee. Millegipärast tundus tollal just tekstide tõlkimine elektronarvutite abil eriti ahvatlev. Esialgsed lootused kiirele ja kergele edule luhtusid peagi. Kuid näpp oli kuradile antud. Peagi leiti, et tõlkimine pole kaugeltki ainus ja ka mitte kõige olulisem ülesanne, mille lahendamine nõuab arvutite rakendamist inimkeele analüüsimiseks. Tänapäeval on tekstide automaattöötlus muutunud niivõrd laiaks ja kirjuks tegevusalaks, et pole sugugi lihtne seda üheselt mingitesse lahtritesse jagada.

Kui rääkida kõigepealt tekstide laias mõttes töötlemiseks loodud ja loodavatest süsteemidest, siis võib siin kõnelda kolmest põhisuunast. Esimesena nimetame süsteeme, mille kõige üldisemaks ülesandeks on nõutava informatsiooni otsing algselt keeleliste tekstidena vormistatud ja arvuti mällu viidud massiivist; need on lühidalt automatiseeritud i n f o o t s i s ü s t e e m i d. Siia võib arvata ka süsteemid, mis küll pole loodud otseselt vajaliku informatsiooni otsimiseks, kuid kaudselt seisavad sama eesmärgi teenistuses, näiteks tekstide automaatrefereerimise ja -annoteerimise süsteemid. Kõik need esindavad keele automaattöötluse kõige praktilisema orientatsiooniga suunda. Arvutilingvistikaga on nende süsteemide rajamine seotud vaid osaliselt, sedavõrd kui tegemist on tekstide, seega loomuliku keele töötlemisega. Teisest küljest kuulub infootsisüsteemide loomine informaatikasse — teadusharusse, mis ametliku liigituse järgi kuulub hoopis tehnikateaduste alla ja mille eesmärgiks on informatsiooni (see-

juures mitte ainult tekstide kujul esitatud informatsiooni) kogumise, säilitamise, töötlemise ja otsimise meetodite väljatöötamine.[15]

Tekstide automaattöötluse süsteemide teise olulisema alaliigi moodustavad süsteemid, millest juba oli juttu — *masintõlkesüsteemid*. Tekstide automatiseeritud tõlkimine ühest keelest teise, mis XX saj. 60-ndatel aastatel elas läbi sügava kriisi, on 70-ndatel aastatel taas muutunud elujõuliseks ja aktuaalseks suunaks kompuuterlingvistikas. Masintõlkesüsteemide loomine eeldab läbimõeldud ja detailsete keele kirjeldus-, analüüsi- ja sünteesimeetodite väljatöötamist ning on selles mõttes igati lingvistika alla kuuluv probleem. Loomuliku keele spetsiifika mittearvestamine (ja mittetundmine) ning püüd lahendada probleemi jõumeetodil oligi õigupoolest masintõlke omaaegse kriisi põhjuseks.

Kolmas liik on süsteemid, mille ülesanne on tagada suhtlemine arvutiga loomulikus keeles. Need süsteemid moodustavad nn. *dialoogsüsteemide* eriliigi.[16]

Kuivõrd ehtne suhtlemine eeldab, et arvuti «mõistaks» nende keeleväljendite sisu, millega inimene tema poole pöördub, siis nimetatakse neid ka keelt mõistvateks süsteemideks. Keeleoskus ja sellele tuginev teksti sisu mõistmine on üks inimintellekti spetsiifilisemaid ja keerulisemaid avaldusi. Seetõttu tuntakse keelt mõistvate süsteemide vastu peale arvutilingvistika huvi veel ühes noores, kuid väga intensiivselt arenevas teadusharus — *tehisintellektis*. Selle omapärase nimega on hakatud nimetama küberneetilist distsipliini, mille eesmärk on meetodite ja süsteemide väljatöötamine ülesannete lahendamiseks, mis nõuavad lahendajalt küllalt kõrget intellekti, alates näiteks malemängust või teoreemide tõestamisest ja lõpetades haigele diagnoosi panekuga mingi aparaadi abil või ehitise projekteerimisega etteantud tingimustest lähtudes; ka robotite loomine kuulub tehisintellekti alla. Nimetatud ülesanded on intellektuaalse infotöötluse näited, mille automatiseerimise vajadusest oli juttu esimeses jaotises. Keelt mõistvate süsteemide rajamine kuulub siia kahel põhjusel: esiteks kujutab teksti mõistmine endast intellektuaalset infotöötlust kõige ehtsamas mõttes, teiseks aga on loomulik nõuda, et sellise süsteemiga saaks ka suhelda intellektuaalselt, s. o. loomulikus keeles.

Need kolm liiki keele automaat-

---

[15] Informaatikaga ja eriti infosüsteemidega informaatika vaatevinklist vaadatuna võib tutvuda raamatus *Informaatika ABC*. Toim. U. Agur. Tln., 1980.

[16] Arvutustehnikas nimetatakse dialoogsüsteemideks üldiselt sääraseid süsteeme, milles ülesande lahenduse leidmine (olgu see arvutusülesanne, informatsiooni otsing, ehitise või aparaadi projekteerimine vms.) kulgeb individuaalpuldi taga töötava inimese ning arvuti kahekõnena, vastastikku küsimusi esitades ning repliike vahetades. Dialoog toimub loomuliku keele sõnu, teatud standardkonstruktsioone ja -väljendeid kasutades. (*Toim.*)

töötluse süsteeme esindavad otsekui järjestikuseid astmeid teel inimkeele üha täielikumale mõistmisele arvutite poolt. Probleemid, mis sellel teel tuleb lahendada, ja selleks kasutatavad meetodid, moodustavad järgnevate peatükkide põhiteema.

Keele automaattöötlus ei ammendu kirjeldatud tüüpi süsteemide loomisega. Lisaks sellele kasutatakse arvuteid tänapäeval keelematerjali töötlemiseks väga mitmesugustel muudel eesmärkidel: tekstide statistiliseks analüüsiks, mitmesuguste sõnastike, sealhulgas sagedussõnastike ja tesauruste koostamiseks, aga ka hääldatava sidusa kõne analüüsiks ja sünteesiks. Et need kompuuterlingvistika spetsiifilisemad rakendused jäävad meie valitud põhiteemast kõrvale, siis me nendel raamatus lähemalt ei peatu.

Mõni sõna veel arvutuslingvistika olukorrast Eesti NSV-s, kuivõrd raamatus ka sellest eraldi juttu ei tule.

Arvutilingvistika-alaste tööde alguse võib meil paigutada 1960-ndate aastate algupoolele, kui TRÜ matemaatiku ning TRÜ arvutuskeskuse rajamise algataja Ü. Kaasiku eestvõttel püüti korraldada vene-eesti masintõlketöid. Kuivõrd peagi jõuti veendumusele, et tegelikuks tööks läheb vaja hoopis ulatuslikumaid keelealaseid uuringuid kui alguses kujutletud (sellest tuleb lähemalt juttu 3. peatükis), siis jäi töö paratamatult soiku.

Tänapäeval arenevad mainimist väärivad arvutilingvistika-alased tööd Eestis kahes kohas — ENSV TA Keele ja Kirjanduse Instituudi arvutuslingvistika sektoris ja Tartu Riiklikus Ülikoolis, kus nendega tegeldakse juba pikemat aega tehisintellekti probleemgrupis ning keelestatistika uurimisrühmas.

KKI arvutuslingvistika sektor on spetsialiseerunud keelealase materjali automaattöötlusele keeleteaduslikel eesmärkidel. Arvutisse on viidud kõigile hästi tuntud «Õigekeelsussõnaraamat»; soovi korral võib sellest automaatsõnastikust saada kiiresti andmeid ühtede või teiste morfoloogiliste tunnustega sõnade kohta, näiteks lasta arvutil välja trükkida kõik eesti keele tegusõnad või ainult kindlaid tunnuseid omavad tegusõnad jne.

TRÜ keelestatistika uurimisrühm on viimastel aastatel intensiivselt tegelnud eesti keele sagedussõnastiku koostamisega. Seni on avaldatud materjale eesti ilukirjanduse sagedussõnastikust TRÜ väljaandel ilmuvais kogumikes «Keelestatistika».

TRÜ tehisintellekti probleemgrupis on tegeldud automatiseeritud infootsisüsteemide probleemidega. On välja töötatud automatiseeritud süsteem JURIOS eestikeelsete seadusandlike tekstide otsimiseks, mis aga alles ootab realiseerimist. Viimastel aastatel on grupp asunud ka keelt mõistvate dialoogsüsteemide kallale, tegeldes eelkõige vajalike keele kirjeldamise meetodite ning mõistmisprotsessi mudeli väljatöötamisega.

2. Tekstide otsing ja töötlus automatiseeritud infosüsteemides

Erinevalt inimese ajust ei unusta arvuti kunagi, mida talle on öeldud, ning võib vastava signaali saamisel kõik jälle välja laduda. Seepärast ongi ehk kõige kasulikum arvutile üldse mitte midagi ütelda.

C. Ford, «Mõtlemise õpetus»

2.1. Automatiseeritud infootsisüsteemid — probleemid ja ehitus

Vajamineva informatsiooni ülesotsimine kui omaette probleem sündis ilmselt ühel ajal esimeste arhiivide ja raamatukogude tekkega, seega 4—5 aastatuhandet tagasi. Juba kolme ja poole tuhande aasta eest valitsenud vaarao Ramses II raamatukogu sisaldanud 20 000 papüürusrulli, nii et vajaliku informatsiooni leidmine ei saanud tollalgi olla triviaalne ülesanne.

Möödaläinud aastatuhandete jooksul on välja töötatud küllalt tõhusad meetodid ja vahendid vajalike raamatute, ajakirjade ja artiklite ülesotsimiseks. Kellel on olnud vähegi kokkupuutumist raamatukogudega, see tunneb hästi põhilisi neist vahendeist — need on raamatute ja ajakirjade kataloogid, artiklite kartoteegid, samuti bibliograafilised teatmeteosed.

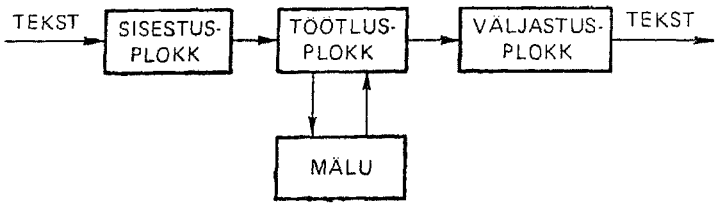
Kuid eelmises peatükis toodud faktid tõendavad, et need vahendid ei suuda tänapäeva tohutute informatsioonimassiivide juures enam kaugeltki rahuldada informatsiooni tarbijate üha pakilisemaid ja spetsiifilisemaid vajadusi. Meie eesmärgiks ei ole muidugi raamatukogude töötamispõhimõtete kirjeldamine ega ka neis seni kasutusel olnud infootsingumeetodite analüüs. Kuid raamatukogude kui näite varal jõuame kergemini meid huvitavate probleemide juurde.

Kui meid huvitab mingi kindla küsimuse kohta leiduv kirjandus, siis peamine allikas selle väljaselgitamisel raamatukogus on süstemaatiline kataloog. Selles on kogu raamatukogus leiduv kirjandus klassifitseeritud sisu järgi kindlatesse liikidesse, mis on hierarhilise ehitusega, s. t. jagunevad järjest kitsamateks alaliikideks. Tänapäeval on enamikus maailma maades kasutusel ühtne teadus- ja tehnikakirjanduse liigitussüsteem: kümnendliigitus ehk rahvusvahelise nimetusega universaalne detsimaalklassifikatsioon, lühendatult UDK. Selles on teadusalad jagatud kõigepealt kümneks pealiigiks, iga pealiik edasi kümneks liigiks, iga liik kümneks alaliigiks jne. Kui meid huvitab mingi konkreetne probleem, siis peame kõigepealt kindlaks tegema, missuguse alaliigi alla ta kuulub, ning seejärel

võime kataloogist leida selle alaliigi kohta raamatukogus leiduva kirjanduse.

Sellest iseloomustusest selguvad ka kirjeldatud meetodi põhilised puudused. Kõigepealt on teemade liigitus üpris jäigalt ette antud. Mingist konkreetsest probleemist lähtudes ei ole alati sugugi lihtne otsustada, missuguses UDK alaliigis võib paikneda teda käsitlev kirjandus. Pealegi asetsevad paljud probleemid mitme teadusharu piirimail. Teiseks peab süstemaatilise kataloogi klassifikatsioon paratamatult peatuma mingite küllalt üldiste allteemade tasandil ning teda ei ole mõeldav lõpmatuseni edasi arendada. Tegelikud probleemid, mille kohta spetsialistil võib informatsiooni vaja minna, võivad olla lõpmata mitmekesised ja puudutada kui tahes kitsast lõiku vastavast teadusharust. Nende kohta süstemaatilisest kataloogist kirjanduse leidmiseks peab ikka läbi vaatama kataloogi kogu vastava alajaotuse või alajaotused. Ja lõpuks, kõnealust tüüpi kataloogist või kartoteegist leiame üldreeglina ainult väljaannete bibliograafilised kirjed (autorid, pealkirja jms.), paremal juhul koos mõnelauselise sisukirjeldusega. Pealkiri ja annotatsioon suudavad aga harva avada raamatu või ka artikli sisu niisugusel määral, et spetsialist suudaks otsustada, kas tema spetsiifilist probleemi selles on käsitletud või mitte, see tähendab, kas tal tasub seda raamatut või artiklit lugeda või mitte. Anda kataloogis või kartoteegis iga raamatu või artikli ammendav sisukirjeldus — selleks ei ole raamatukogud praeguse töökorralduse juures füüsiliselt suutelised.

Siit jõuame aga küsimuse juurde, kas raamatukogu ülesanne ongi üldse anda informatsiooni sellest, mida ühe või teise probleemi kohta on avaldatud. Raamatukogu algne funktsioon oli ju muu: olla raamatute, ajakirjade jms. säilituskohaks ning laenutada neid inimestele, kes juba teavad, mida nad tahavad. Teenindada inimest, kes tuleb otsima mingit kindlat trükist, on hoopis midagi muud kui teenindada inimest, kes tuleb erialase probleemiga ja tahab teada, kas ja kus selle kohta on midagi kirjutatud. Seda enam, et viimasel juhul ei huvita inimest sageli mitte niivõrd see, mida on leida antud raamatukogus, vaid maailmas üldse. Nende kahe ülesande lahendamine eeldab erinevaid meetodeid, seda nii teoreetilises kui ka organisatsioonilises mõttes. Sedamööda, kuidas sai selgemaks nende ülesannete põhimõtteline erinevus, kujunes raamatukogusüsteemi kõrval teine teenindussüsteem, mida nimetatakse infoteeninduseks. Selles ei säilitata ega väljastata mitte niivõrd konkreetseid dokumente, kuivõrd informatsiooni nende sisust. Ühtlasi kujunes raamatukogunduse kui teoreetilise distsipliini ja mitme muu teaduse najal uus iseseisev teadusharu i n f o r m a a t i k a. Raamatukogule vastab infoteeninduses hoopis teist laadi organisatsiooniline üksus — infosüsteem, mille keskme moodustab i n f o r m a t s i o o n i o t s i m i s e s ü s t e e m ehk lühidalt i n f o o t s i-



Joon. 2

süsteem. Infootsisüsteemide loomisega ja nende töömeetodite väljatöötamisega seotud probleemid moodustavad informaatika põhilise uurimisobjekti.[17]

Meidki huvitavad käesolevas raamatus infosüsteemid, kuid ainult nende üks liik — automatiseeritud süsteemid, mis on realiseeritud arvutitel. Needki esmajoones ühest aspektist: meid huvitab, missugused tekstitöötlusprotsessid leiavad aset neis süsteemides ning missuguseid meetodeid kasutatakse tekstide töötlemiseks ja nende sisu esitamiseks.

Mida kujutab endast automatiseeritud infootsisüsteem? Hiljem kirjeldame tema ehitust üksikasjalikumalt, praegu aga piisab tööta-

[17] Informaatika ja raamatukogunduse vahele ei ole siiski võimalik täiesti kindlat piiri tõmmata, nii nagu ei saa vastandada ka raamatukogusid ja infotalitusi. Näiteks kujutavad süstemaatilised ja muud kataloogid, millest eespool rääkisime, endast tegelikult raamatukogusiseseid infootsisüsteeme ning neid võibki pidada tänapäeva infootsisüsteemide eeskujudeks. Kuid küsimuse see aspekt praegu meid ei huvita.

mispõhimõtete selgitamiseks järgmisest lihtsast skeemist (joon. 2).

Infootsisüsteemi keskseks komponendiks on mälu; siin säilitataksegi informatsiooni, mille kohta süsteemile võib esitada küsimusi ja mille ulatuses ta on suuteline vastama. Informatsioon oli algselt esitatud tekstidena, mida informaatika terminoloogias nimetatakse lihtsalt dokumentideks, olgu need siis teaduslikud artiklid, aruanded, tehnoloogiliste protsesside kirjeldused vm. Sisestusploki ülesandeks on teisendada esitatavad tekstid süsteemisisesele kujule, s. o. sellisele kujule, millega süsteemi kandjaks olev arvuti saab opereerida. Tekst võib kujutada endast süsteemi mällu viidavat järjekordset dokumenti, ent samuti süsteemile esitatud küsimust ehk päringut. Vastuseks päringule peab süsteem otsima mälust tekstid — dokumendid —, mida päring puudutab. Niihästi uute tekstide mälluviimist kui ka vastuseks sobivate tekstide otsimist mälust korraldab töötlusplokk. Väljastusplokis teisendatakse vastusena väljastatav informatsioon sellisele kujule, millega ini-

mesel on mugav opereerida, s. o. tekstideks.

Vastuse otsimine seisneb selles, et süsteem võrdleb talle esitatud päringut kõigi mälus leiduvate tekstide teatud tunnustega ning teeb kindlaks, missugustes tekstides käsitletakse päringus nimetatud probleeme. Tänu elektronarvuti tohutule töökiirusele kulub võrdlemiseks ainult mõni sekund või minut isegi sel juhul, kui süsteemi mälu sisaldab kümneid tuhandeid dokumente. Lihtsamal juhul võib võrdlemine seisneda selles, et arvuti kontrollib, missugustes tekstides sisalduvad päringus kasutatavad sõnad. Niisugune triviaalne kõrvutamine pole sageli küllalt tõhus, sest päringus leiduvate sõnade esinemine tekstis ei taga kaugeltki päringule s i s u l i s e l t vastavate tekstide äratundmist. Tasub vaid meelde tuletada eelmises peatükis kirjeldatud loomuliku keele väljendite iseärasusi — suur varieeruvus (sünonüümia), homonüümia, elliptilisus. Selleks et otsing oleks efektiivsem, tõlgitakse nii süsteemi viidavad dokumendid kui ka päringud loomulikust keelest erilisse i n f o k e e l d e ning süsteemi sees opereeritakse ainult infokeele väljenditega. Infokeeled on formaliseeritud tehiskeeled, mis on loodud kasutamiseks infootsisüsteemides. Infokeel saadakse loomuliku keele, täpsemini, vastava eriala allkeele erilise töötluse tulemusena. Selles ei opereerita enam sõnade, vaid mõistetega, mille jaoks võidakse võtta kasutusele omad tähistused. Infokeele põhiülesanneteks on neutraliseerida loomulike keelte varieeruvus, tagada, et igale mõistele vastaks ainult üks väljend ja igale väljendile vaid üks mõiste, samuti fikseerida keelde kuuluvate mõistete vahel kindel loogiliste vahekordade süsteem.

Infokeeli on väga mitmesuguseid, laiemalt tuntud ja kasutatavad on aga kaks põhiliiki — liigituskeeled ehk klassifikatsioonkeeled ning deskriptorkeeled.

L i i g i t u s k e e l on mingi teadus- või tegevusala või ka kogu inimteadmuse hierarhiline liigitus järjest kitsamateks üksteist välistavateks alaliikideks, kus iga alaliiki tähistab kindel sümbol. Püüd välja töötada inimteadmisi süstematiseerivaid liigitusi on olnud inimestele omane juba ammu, Aristotelese kategooriate süsteem on selle idee üks teostusnäiteid. Need klassifikatsioonid teenisid aga rohkem filosoofilisi eesmärke kui infootsingut ja muid praktilisi ülesandeid. Tänapäeval on tuntuim ja seejuures nimelt infokeelena kasutatav liigitus *universaalne detsimaalklassifikatsioon* (UDK) ehk kümnendliigitus, mille põhimõtted töötas välja möödunud sajandil ameeriklane M. Dewey. Käesoleva sajandi alguskümnenditel täiustas seda Rahvusvaheline Bibliograafia Instituut ning 30-ndatest aastatest alates on see kasutusel juba paljudes maades. Nimetus *«kümnendliigitus»* tuleneb asjaolust, et selles jaotatakse teadusalad, nende liigid ja alaliigid järjekindlalt kümneks väiksemaks alaliigiks. Kümme kõige laiemat pealiiki on järgmised:

nim. ENSV Riiklik Raamatukogu

0. Üldosa. Kultuur. Haridus.
1. Filosoofia.
2. Religioon.
3. Ühiskonnateadused.
4. Filoloogia.
5. Matemaatika. Loodusteadused.
6. Rakendusteadused.
7. Kunst. Sport.
8. Ilukirjandus.
9. Ajalugu. Geograafia.

Liikide ja alaliikide tähistamiseks kasutatakse sõnaliste väljendite asemel numbrilist kodeerimisviisi kui universaalset (konkreetsest keelest sõltumatut) ja pealegi märksa lühemat. Alaliikide numberkoode nimetatakse vastavate mõistete indeksiteks. Näiteks mõistele *vahelduvvoolumasinad* vastab indeks 621.313.3 (6 — rakendusteadused, 62 — tehnika, 621 — üldine masinaehitus, 621.3 — elektrotehnika, 621.31 — elektroenergeetika, 621.313 — elektrimasinad, 621.313.3 — vahelduvvoolumasinad). Niisiis sisaldab iga mõiste indeks ühtlasi informatsiooni kogu mõistehierarhia kohta, millele antud mõiste allub.

Lisaks niisuguste hierarhiate esitamisele on UDK-s välja töötatud vahendeid mittehierarhiliste vahekordade, samuti vastavaid mõisteid puudutavate täienduste ja täpsustuste esitamiseks. Näiteks kirjeldab indeks [331.2:621.311 (~+47.31) «1917:23»]=2 järgmist sisu: «statistilised andmed elektrijaamade töötajate töötasu kohta kogu maailmas võrrelduna Moskva oblastiga aastatel 1917—1923, kirjutatud inglise keeles.» Nagu näeme, võib UDK abil kirja panna tervete referaatide sisu. Ta on seega ehtne formaalne keel ja mitte lihtsalt mõistete hierarhilise süsteemi kirjeldusvahend.

Peale UDK on kasutusel ka teisi liigituskeeli, niihästi universaalseid nagu UDK kui ka mõne kitsama ainevalla jaoks looduid. Raamatukogunduses on sellised infokeeled osutunud üpris tõhusaiks; eks seda tõenda asjaolu, et UDK-d kasutatakse tänapäeval raamatukogudes üle terve maailma.

Automatiseeritud infootsisüsteemides kasutatakse liigituskeeli seni siiski võrdlemisi vähe. Põhjus on järgnevas. Kui dokumendi sisu on juba tõlgitud liigituskeelde, siis suudab süsteem vastavate indeksväljenditega küll edukalt opereerida; probleemiks on aga loomulikus keeles kirjutatud teksti tõlkimine liigituskeelde. Teksti tõlkimist infokeelde nimetatakse üldiselt tema indekseerimiseks. Raskusi tekitab nimelt teksti automaatne indekseerimine: protsess, kus arvuti «loeb» teksti ja esitab selle sisu UDK või muu liigituskeele indeksina või indeksitena. Indeks ei seostu otseselt dokumendi teksti sõnastusega, vaid ainult selle sisuga ja sedagi kaudselt, mitmete vahendusoperatsioonide kaudu. Kui inimene indekseerib mingit dokumenti, näiteks teaduslikku artiklit, siis peab ta selle kõigepealt läbi lugema, artikli sisust aru saama, seejärel koostama sisu kokkusurutud, kuid võimalikult täpse kirjelduse ja alles seejärel saab ta moodustada artikli indeksi. Nagu hiljem näeme, osu-

tuks kirjeldatud protsessi täies ulatuses automatiseerimine praegu üle jõu käivaks ülesandeks isegi kõige täiuslikumate keelt mõistvate eksperimentaalsüsteemide puhul, rääkimata siis praktiliseks ja massiliseks kasutamiseks mõeldud infootsisüsteemidest. Seetõttu tuleb infootsisüsteemides, kus informatsiooni süsteemisisese esitamise vahendina kasutatakse liigituskeeli, tekstide indekseerimine paratamatult jätta inimeste teha. Niisugune käsitsitöö etapp vähendab aga tunduvalt automatiseeritud süsteemi efektiivsust, seda eriti süsteemide puhul, kus mällu viidavate tekstide hulk on suur või kus seda tuleb pidevalt uuendada ja täiendada.

Liigituskeeltel on veel teisigi puudusi, mis raskendavad nende kasutamist isegi sel juhul, kui tekstid indekseeritakse käsitsi. Nii eeldavad nad, et tekstis käsitletavad probleemid oleksid üheselt ja selgelt kindlatesse kategooriatesse liigitatavad. Spetsiifiliste ja mitmetahuliste probleemide kirjeldamine (vrd. eespool toodud näidet) nõuab ka inimeselt üpris suurt intellektuaalset pingutust ning sellega saavad hakkama vaid kvalifitseeritud spetsialistid.

Kõige selle tõttu ongi liigituskeeled infokeeltena automatiseeritud infootsisüsteemides vähe levinud. Peamiseks infokeelte tüübiks on neis deskriptorkeeled. Deskriptorid on keele üksused, mis esitavad vastava ala kindlaid mõisteid — umbes nagu loomuliku keele sõnad nimetavad reaalseid nähtusi, millest inimestel on vaja kõnelda. Vormiliselt ongi deskriptorid sõnad või sõnaühendid, mis erinevalt liigituskeele üksustest on valdavalt võetud otse vastava eriala tekstidest. Seetõttu seisneb deskriptorite puhul teksti indekseerimine deskriptoritele vastavate sõnade ja sõnaühendite ülesotsimises tekstist. See protsess on praegu arvutile juba täiesti jõukohane.

Deskriptorkeelega lahutamatult seotud mõiste on tesaurus — deskriptorite erilisel viisil korrastatud sõnastik.

Et tesaurused on nüüdisaegsete automatiseeritud infootsisüsteemide keskseid lülisid ning et just tesauruste koostamise ja kasutamisega on siin seotud põhilised keelelise analüüsi probleemid, siis peatume neil veidi lähemalt.

Tesaurusi kui eriliiki sõnastikke on koostatud ammu enne seda, kui hakati mõtlema automatiseeritud infootsisüsteemidele. Tesaurust selle laias tähenduses nimetatakse ka *ideograafiliseks sõnastikuks.* See on sõnastik, kus on ühte koondatud ühe kindla mõiste või ideega seotud sõnad ja väljendid; see tähendab, sõnade korrastamise aluseks sõnastikus ei ole nende vorm (nagu tavalistes alfabeetiliselt korrastatud sõnastikes), vaid sisu. Juba II sajandist m. a. j. on tuntud Julius Polluxi poolt kreeka keele baasil koostatud «Onomastikon». See koosneb 10 raamatust, millest igaüks haarab teatud kindla teemaga seotud sõnu. Esimene raamat sisaldab näiteks sõnu, mis on seotud jumalate ja kuningatega, seitsmes — kaubandusega seotud sõnavara jne. Üldkeelelistest te-

saurustest tuntuim on kindlasti inglase M. R. Roget' poolt koostatud «Thesaurus of English Words and Phrases», mis esmakordselt ilmus 1852. aastal. Pisut täiendatud kujul on see kasutusel tänapäevalgi. Roget jagab kogu keele sõnavara kõigepealt nelja üldisesse mõisteklassi: abstraktsed vahekorrad, ruum, mateeria ja vaim (mõistus, tahe, emotsioonid). Need mõisteklassid jagunevad kategooriateks, kategooriad sektsioonideks, sektsioonid mõisterühmadeks. Kokku on tesauruses 1000 mõisterühma, mille vahel jaotuvad sõnad.

Infootsisüsteemides kasutatav tesaurus on pärinud nii oma nime kui ka esialgsed materjali korrastuspõhimõtted üldkeelelistelt sõnaraamatutelt, kuid tema ehitus on hoopis rangemalt kindlaks määratud. Mitmel maal, sealhulgas ka Nõukogude Liidus on ametlikult kehtestatud infootsitesauruste koostamise ja vormistamise riiklikud standardid[18], samuti on UNESCO esitanud juhendid üheja ka mitmekeelsete tesauruste koostamiseks.

Eeltoodud kirjeldustest võib jääda mulje, et tesauruste ja liigitussüsteemide vahel pole siiski eriti suurt põhimõttelist vahet: ka tesaurus klassifitseerib ju temas sisalduva materjali kindlatesse rubriikidesse, mis on hierarhiliselt korrastatud. Kuid nagu juba osutasime, on infootsitesaurustel põhinevate infokeelte ja liigituskeelte vahel siiski see põhimõtteline erinevus, et tesauruses moodustavad korrastatava materjali sõnad ja sõnaühendid, mis on võetud selle ainevalla tekstidest, mille jaoks tesaurus — ja otsisüsteem tervikuna — on loodud. Tesauruse kogu ülejäänud organisatsioon teenib vaid üht eesmärki: osutada sisulisi (semantilisi) s e o s e i d mõistete vahel, mida tähistavad tesauruses sisalduvad sõnad ja sõnaühendid. Infootsitesauruse kui terviku loogiline struktuur kujunebki suurel määral nende seoste määramise tulemusena, teda ei anta aprioorselt ette nii nagu liigituskeelte puhul. Küll on aga tesauruses kajastada tulev mõistetevaheliste seoste ring võrdlemisi kindlalt ette määratud. Vaatlemegi neid seoseid pisut lähemalt, et saaks selgemaks ka tesauruse üldine ehitus ja tema roll automatiseeritud infootsisüsteemis.

Eristatakse kolme põhiliiki semantilisi seoseid sõnade ja mõistete vahel tesauruses: ekvivalentsusseos (sünonüümia ja kvaasisünonüümia), hierarhilised seosed (sooliigiseos) ning assotsiatiivseosed.

E k v i v a l e n t s u s s e o s kehtestatakse sõnade ja sõnaühendite vahel, millega väljendatavad tähendused (mõisted) loetakse antud informatsioonisüsteemi raames samaväärseteks. Sellisteks sõnadeks on kõigepealt sünonüümid — sõnad, mis omavad sama tähendust ning tähistavad üht ja sama mõistet: *keeleteadus* ja *lingvistika, rahandusplaan* ja *finantsplaan, organiseerimiskoosolek* ja *asutamiskoosolek*. Vajadus niisugused väl-

[18] ГОСТ 18383—73. Тезаурус информационно-поисковый. Общие положения. Форма представления. Введ. 01. 07. 1973 г. М., 1974.

jendid ühte koondada on arusaadav: kui otsisüsteemile esitatakse küsimus mingi nähtuse või sündmuse kohta, siis on vaja, et ta otsiks üles kõik tekstid, milles antud nähtusest või sündmusest on räägitud, ükskõik missuguse väljendi abil seda nimetatakse. Näiteks võime seadusandlikes tekstides leida väljendeid *joobnud, purjus, joobeseisundis, alkoholijoobes, ebakaines olekus,* mis kõik tähistavad üht ja sedasama seisundit. Kui soovime kätte saada tekste, milles on juttu kõnealusest seisundist, siis tuleb otsida kõiki dokumente, kus ükskõik missugune neist sünonüümidest ette tuleb.

Kuid lisaks ehtsatele sünonüümidele loetakse infootsitesaurustes samaväärseteks ka paljud sõnad ja muud väljendid, millel pole küll täpselt sama tähendus, kuid mida antud otsisüsteemi raames ei ole mõtet eristada: kas on vastav tähenduslik erinevus liiga ebaoluline või siis sisaldub tekstis, milles räägitakse ühest mõistest, alati informatsiooni ka teise mõiste kohta. Niisuguseid väljendeid nimetatakse tinglikeks ehk *kvaasisünonüümideks* (*kvaasi* 'peaaegu', 'pool-'). Erinevatel eesmärkidel loodud otsisüsteemides võivad ühel ja samal sõnal olla muidugi hoopis erinevad kvaasisünonüümid, seetõttu saavad näitedki siin olla vaid tinglikud. Nii võib majandus- või haldustekstides lugeda ekvivalentseteks väljendid *tõstmine, kasv, tõus, suurenemine,* samuti *aeg, tähtaeg, periood,* tehnikaalastes tekstides *orbiit, apogee, perigee.* Mõnikord on otstarbekas lugeda kvaasisünonüümideks isegi antonüümid, s. o. vastandliku tähendusega sõnad, sest just tähenduste vastandlikkuse tõttu on sellised sõnad üksteisega lahutamatult seotud.

Ekvivalentsusseose kehtestamine väljendite vahel on üks peamisi vahendeid loomuliku keele varieeruvuse neutraliseerimiseks. Selles seoses olevad väljendid tähistavad kõik üht ja sama mõistet või kvaasimõistet. See mõiste tähistatakse kindla väljendiga, milleks reeglina valitakse üks eelistatavaim ekvivalentsusseoses olevatest sõnadest või sõnaühenditest. Deskriptori all mõeldaksegi rangelt võttes vaid mõistet koos tema tähistajaks valitud väljendiga. Teised sõnad ja sõnaühendid, mis samasse ekvivalentsusklassi kuuluvad, ei ole deskriptorid, vaid lihtsalt märksõnad, antud deskriptori sünonüümid. Ülejäänud seostes, mida tesaurus kajastab, osalevad ainult deskriptorid.

Muudest semantilistest seostest on olulisimad hierarhilised seosed, eriti soo-liigiseos. Üldreeglina kuulub iga deskriptor tesauruses mingisse soo- ja liigiseoste hierarhiasse, harilikult isegi mitmesse eri hierarhiasse (muidugi erinevate liigitusaluste järgi). Võtame näiteks sellise kõigile tuttava mõiste nagu *töötasu.* Sellel võime leida erinevatest liigitusalustest lähtudes mitmeid alaliike, s. o. liigimõisteid: *palk, kuutöötasu, ületunnitasu, lisatöötasu* jne. Töötasu mõiste ise on aga liigimõisteks sellise mõiste suhtes nagu *sissetulek:* töötasu on üks liik sissetule-

kuid kõrvuti selliste liikidega, nagu *preemia, honorar, stipendium, pension* jt. Mõned teised näited mõistehierarhiatest: *karistus* (juriidilises mõttes) on üks liik *ühiskondlikke mõjustusvahendeid, kriminaalkaristus* — üks liik karistusi, *vabadusekaotus* — üks liik kriminaalkaristusi; *diiselmootor* on üks liik sisepõlemismootoreid, *sisepõlemismootor* — üks liik *mootoreid* jne.

Soo-liigiseose tähtsus informatsiooni otsimisel tuleneb asjaolust, et sisuliselt üht ja sama nähtust, sündmust või olukorda võidakse loomulikus keeles kirjeldada kord üldisema, kord konkreetsema tähendusega sõnade abil. Pealegi, kuivõrd soomõiste haarab mahuliselt kõiki oma liigimõisteid, siis käib iga dokument, mis puudutab soomõistet, automaatselt ka kõigi selle liigimõistete kohta. Kui meid huvitab, kuidas on reguleeritud ühe või teise kategooria töötajate töötasud, siis tuleb silmas pidada, et meile vajalikud andmed võivad sisalduda ühelt poolt dokumentides, mis reguleerivad nende töötajate sissetulekuid üldse, teiselt poolt aga ka dokumentides, mis reguleerivad nende üht või teist töötasu liiki (palka, lisatöötasu, ületunnitasu jne.). Kui vastavad seosed on tesauruses fikseeritud, võib otsisüsteem neid otsingul automaatselt arvesse võtta.

Kolmas infootsitesaurustes regulaarselt kajastatav deskriptoritevaheliste seoste tüüp on a s s o t s i a t i i v s e o s e d. Erinevalt kahest eelmisest seoseliigist ei ole assotsiatiivseos loogiliselt määratletud, vaid ta kehtestatakse nende deskriptorite vahel, millega vastavuses olevad esemed või nähtused on mingil olulisel viisil omavahel seotud, ehkki selle seose olemust pole alati lihtne määratleda. «Olulisel viisil seotud» tähendab siin esmajoones seda, et vastav seos võib meid informatsiooni otsingul juhtida vajalike andmete juurde. Näiteks kui meid huvitavad töötervishoidu käsitlevad normatiivaktid, siis vastavat materjali võime küllalt suure tõenäosusega leida ka tööseadusandlusalastest tekstidest, kus räägitakse nähtustest nagu tolm, ventilatsioon, vibratsioon, tervist kahjustav töö jms. Viimased pole ei töötervishoiuga ekvivalentsed mõisted ega ka selle liigid, kuid nende seos töötervishoiuga on ilmne. Seetõttu on otsarbekas tesauruses mõiste *töötervishoid* siduda ka viimati loetletud mõistetega, kuigi vastavate seoste sisulist olemust ei ole lihtne kindla kategooria alla viia.

Eri ainevaldkondade puhul tuleb oluliseks lugeda ilmselt erinevad seosetüübid. Näiteks tehnikatekstides, kus suure osa informatsioonist (seejuures just sageli nõutavast informatsioonist) moodustavad ehitiste, masinate või seadeldiste kirjeldused, on üheks olulisemaks assotsiatiivseose alaliigiks tervikuosaseos: *hoone — vundament, sein, katus* jne; *sisepõlemismootor — kolb, keps, silinder* jne. Paljudes tehnikatesaurustes on see seos seetõttu esile toodud iseseisva seoseliigina. Teiselt poolt aga näiteks ühiskonnateaduste-alase informatsiooni puhul, kus valdavalt on te-

gemist abstraktsemate nähtuste ja protsessidega, mängib see seos suhteliselt väikest rolli. Hoopis olulisemad on siin mitmesugused funktsionaalse iseloomuga seosed: nähtuste põhjuslik seos, mingi tegevuse puhul selle eesmärk, vahendid, tagajärjed.

Kirjeldatud kolm seosetüüpi esindavad kõige tüüpilisemaid seosetüüpe tesaurustes, mis on mõeldud kasutamiseks automatiseeritud infootsisüsteemides; neid nõuab ka eespool viidatud GOST. On loodud ka märksa mitmekesisema semantiliste seoste arsenaliga tesaurusi. Mida kitsama ainevalla jaoks tesaurus ja otsisüsteem on mõeldud, seda spetsiifilisemaid seoseid saab selles kasutada. Kuid automatiseeritud infootsingu praktika on näidanud, et paljude erinevate ja väga spetsiifiliste seoste kasutamine harilikult ei õigusta ennast, sest komplitseerib liialt nii tekstide indekseerimist kui ka otsimist. Kui näiteks soo- ja liigiseoste puhul on nende roll otsinguprotsessis võrdlemisi selge ning seetõttu ka lihtne otsustada, millal neid kasutada või mitte kasutada, siis spetsiifiliste seoste puhul ei ole see enamasti sugugi nii. Tekib oht sattuda vastuollu automatiseeritud infootsisüsteemide rajamise seniste aluspõhimõtetega — luua süsteemid, mis oleksid võimalikult lihtsad ja samal ajal tõhusad. Ainult vastavale ainevaldkonnale iseloomulike seoste kasutamine viib meid klassikaliste otsisüsteemide juurest keerulisema tüübi — faktiotsisüsteemide juurde, millest tuleb juttu allpool.

Põhilise osa igas tesauruses moodustab *deskriptorite leksikosemantiline loend.* Selles on tähestikulises järjestuses esitatud deskriptorid, iga deskriptori juures aga on toodud esiteks tema sünonüümid ja kvaasisünonüümid ning teiseks kõik deskriptorid, mis on seotud antud deskriptoriga mingi semantilise seose kaudu: nii deskriptori soo- ning liigimõisteid väljendavad kui ka assotsiatiivseoses olevad deskriptorid. Toome paar näidet sellisest esitusest. Esimene näide on võetud Tartu Riikliku Ülikooli õigus- ja majandusinformatsiooni rühma poolt koostatud tesaurusest, teised «Teadus- ja tehnikaterminite tesaurusest», mis on Nõukogude Liidus koostatuist üks ulatuslikumaid.[19] *S* osutab (kvaasi)sünonüümidele, *SD* — soomõisteid esitavatele deskriptoritele, *LD* — liigimõisteid esitavatele deskriptoritele, *AD* — assotsiatiivselt seostuvatele deskriptoritele.

ERGUTUSABINÕU

S: ergutamisvahend
ergutamise vahend
ergutusvahend
SD: abinõu
LD: aukiri
aunimetuse andmine
auraamatusse kandmine
autahvlile paigutamine
premeerimine
AD: ergutamine
ergutusfond
preemia
tänuavaldus

[19] Тезаурус научно-технических терминов. Под ред. Ю. И. Шемакина. М., 1972.

NAVIGATSIOONISÜSTEEMID

S: aeronavigatsioonisüsteemid
kosmilise navigatsiooni süsteemid
merenavigatsioonisüsteemid
SD: navigatsioonivahendid
LD: astronavigatsioonisüsteemid
inertsiaalsed navigatsioonisüsteemid
kombineeritud navigatsioonisüsteemid
raadionavigatsioonisüsteemid
lähinavigatsioonisüsteemid
kaugnavigatsioonisüsteemid
AD: navigatsioon

KOMBINEERITUD NAVIGATSIOONISÜSTEEMID

SD: navigatsioonisüsteemid
LD: astroinertsiaalsed navigatsioonisüsteemid
raadioastroinertsiaalsed navigatsioonisüsteemid

Tesaurus pole tavaliselt mõeldud ainult infootsisüsteemi funktsionaalse komponendina, vaid ka abivahendina süsteemi kasutajale, et ta päringu sõnastamisel oskaks arvestada, missugused väljendid kuuluvad süsteemi infokeelde ja missugused ei kuulu. Seetõttu on tesaurustes lisaks deskriptorite leksikosemantilisele loendile reeglina veel mitmed täiendavad osad, mis aitavad kasutajal kiiresti leida teda huvitavaid deskriptoreid, eelkõige deskriptorite süstemaatiline register ning hierarhiliste seoste register. Esimeses on deskriptorid liigitatud sisust lähtudes temaatilistesse alajaotustesse, yiimaste sees aga jagatud hierarhiliselt üha kitsamateks rühmadeks. Näiteks on eespool nimetatud «Teadus- ja tehnikaterminite tesauruses» deskriptorid jagatud 33 temaatilisse alajaotusse, mis omakorda jagunevad 170 deskriptorirühmaks. Näiteks alajaotus «Lennundus» jaguneb järgmisteks rühmadeks: 1) lennukikonstruktsioonid, 2) lennuvägi, 3) lennu aerodünaamika ja dünaamika, . 4) aerodünaamilised vormid, 5) lennuväe baseerumine, 6) lennuaparaatide varustus, 7) lennujuhtimissüsteemid, 8) lennuaparaatide tüübid.

Materjali niisugune esitus aitab süsteemi kasutajal päringut täpsemini sõnastada: väga sageli on ju nii, et inimene oskab enam-vähem täpselt otsustada, missugusele kitsamale alale teda huvitav probleem kuulub, aga ei ole kindel küsimuse sobivas sõnastuses. Deskriptorite süstemaatiline register juhibki ta sel juhul nende konkreetsete terminite — deskriptorite juurde, mis antud ala kohta tesauruses leiduvad.

Hierarhiliste seoste registris on esitatud deskriptorite hierarhiad soo-liigiseose järgi. Iga hierarhia tipus on deskriptor, millel tesauruses puudub soomõiste, selle all on loetletud tema vahetud liigimõisted, viimaste all omakorda nende liigimõisted jne. Graafiliselt saab deskriptorite omavahelist paiknemist hierarhias osutada näiteks sellega, et iga deskriptor nihutatakse sõltuvalt tema tasemest teatud arv samme paremale. Näiteks näeb eespool mainitud tesauruses üks deskriptorite hierarhia välja selline:

Stardiseadeldised
- Lennuväe stardiseadeldised
- Laevade stardiseadeldised
- Maapealsed stardiseadeldised
- Liikuvad stardiseadeldised
- Raudteeplatvormidel paiknevad stardiseadeldised
- Iseliikuvad stardiseadeldised
- Statsionaarsed stardiseadeldised
- Maa-alused stardiseadeldised
- Pool-maa-alused stardiseadeldised
- Ballistiliste rakettide stardiseadeldised
- Õhutõrjerakettide stardiseadeldised

Punktide arv deskriptori ees osutab, missugusele tasemele hierarhias deskriptor kuulub. Nagu näha, on toodud hierarhias neli taset.

Kõnealuse registri ülesandeks on aidata süsteemi kasutajal koostada päring parajasti niivõrd üldiste või konkreetsete terminite abil, nagu ta seda kõige otstarbekamaks peab. Nagu praktika on näidanud, eksivad inimesed päringu sõnastamisel väga sageli just sõnastuse tasandi valikul: sõnastavad päringu liiga kitsaste terminite abil (sel juhul jääb neil osa vajalikku informatsiooni saamata) või siis, vastupidi, kasutavad liiga üldisi termineid (ja saavad vastuseks palju üleliigset materjali).

Vaatame nüüd lähemalt, kuidas automatiseeritud infootsisüsteem töötab tervikuna. Et meid huvitab esmajoones lingvistiline aspekt, siis süsteemi ehituse ja töötamise tehnilistel aspektidel me kuigivõrd ei peatu.

Infootsisüsteemi raames esitab tesaurus infokeele sõnavara, mille abil süsteemi mällu viidavad dokumendid — tekstid — indekseeritakse, s. t. tõlgitakse sellisesse vormi, millega arvuti suudab opereerida. Süsteemi viidavat teksti võrreldakse tesauruses leiduvate deskriptoritega ja nopitakse tekstist välja sõnad ning sõnaühendid, millele vastab tesauruses mingi deskriptor, s. o. mis eeldust mööda nimetavad antud ala olulisi mõisteid ning teksti sisu sõlmpunkte. Need deskriptorid osutavad, et antud tekstis on juttu «sellest ja sellest ja sellest...». Niimoodi saadud deskriptoriloend ehk dokumendi *otsitunnustik* jääbki esindama teksti süsteemi mälus.

Kirjeldatud indekseerimisprotsess toimub paljudes süsteemides veel käsitsi, kuid suurte tekstimassiivide puhul on ainumõeldav automaatindekseerimine. Viimasega on seotud mitmed puht keelelised probleemid.[20] Kõigepealt on asi selles, et tundmaks tekstis ära mingit sõna — deskriptorit või selle kvaasisünonüümi — on vaja ära tunda ja deskriptoriga samastada vasta-

[20] Me ei peatu siin ühel puht tehnilisel, aga kaugeltki mitte tühisel probleemil: kuidas tekstid arvutisse viia. Kuni pole automaatseid lugemisseadmeid, mis oleksid suutelised lugema suvalises trükikirjas teksti, on see tegelikult üks esimesi tõsiseid probleeme, millega põrgatakse kokku suurte tekstimassiivide töötlemisel.

va sõna kõik erinevad morfoloogilised avaldumisvormid. Kui deskriptoriks on näiteks mõiste *maksmine*, siis tuleb see ära tunda ka vormides nagu *maksmise, maksmisest, maksmistest* jne., aga ka *maksis, makstud, makstakse* jne., sest tuum on kõigil neil vormidel üks ja sama, kõik nad kuuluvad mõiste *maksmine* alla. Selle probleemi lahendamiseks on kaks teed. Esimene põhineb nii-öelda jõumeetodil ning seisneb selles, et tesauruse «süsteemisiseses» variandis ei esitata deskriptorina vastava sõna täielik kuju, vaid selle lühendatud vorm — tüvi või «kvaasitüvi» sellise arvestusega, et see ühtiks võimalikult paljudel antud sõna morfoloogilistel vormidel. Vaadeldud näite puhul sobib selleks vorm *maksm,* mis on ühine kõigil sõna *maksmine* käänetel, aga ka vormidel *maksma, maksmas, maksmast.* Niisugune lühendatud kuju ei tohi aga ühtida ühegi teise deskriptoriga (tüvega) tesauruses. Oletame näiteks, et kõnealuses tesauruses on ka *maks* (omastav *maksu*) omaette deskriptoriks. Sel juhul ongi *maksm* maksimaalselt lühendatav vorm sõnast *maksmine*, kuid see ei võimalda vaadeldavat deskriptorit ära tunda paljudes vormides, näiteks *maksta, makstakse, maksti, makstud, maksab, maksavad.* Siis ei jää muud üle, kui tesauruse süsteemisiseses variandis lisada kõnealusele deskriptorile veel rida «sünonüüme» vormide näol, mis võimaldaksid ära tunda ka ülejäänud sõnavormid tekstis, näiteks: *makst* (ühendab vorme *maksta, makstud, makstakse*), *maksa* (*maksab, maksad, ei maksa* jne.). Nagu neist näidetest nähtub, ei pruugi niisugusel sõnavormide lühendamisel olla kuigi palju ühist vastava keele sõnade tegeliku morfoloogilise liigendusega, vaid selle määravad puht praktilised kaalutlused.

Teine võimalus seisneb selles, et indekseerimise käigus kasutatakse ühe alloperatsioonina sõnade automaatset morfoloogilist analüüsi. Sel juhul võib tesauruses esitada deskriptorid nende normaalsel kujul, deskriptorite äratundmine tekstis jääb morfoloogilise analüüsi protseduuride kanda, mis peavad suutma kindlaks teha iga sõna kõik morfoloogilised vormid (automaatse morfoloogilise analüüsi põhimõtteid kirjeldame lähemalt allpool). Niisugune lähenemisviis on muidugi palju loomulikum ja seetõttu ka teoreetiliselt eelistatav. Seni on tema rakendamist infootsisüsteemides takistanud asjaolu, et loomuliku keele teksti automaatne morfoloogiline analüüs on võrdlemisi keeruline, eriti võrreldes teiste süsteemis kasutatavate protseduuridega. Sedamööda, kuidas on välja töötatud mitmeid küllalt kiireid ja tõhusaid automaatse morfoloogilise analüüsi algoritme, on neid hakatud viimasel ajal üha laiemalt kasutama ka infootsisüsteemides.

Senine jutt puudutas süsteemi mällu viidavate tekstide indekseerimist. Põhimõtteliselt niisamuti toimub ka süsteemile esitatavate päringute tõlkimine süsteemisisesele kujule, ent siin lisanduvad mõned spetsiifilised probleemid. Kõigepealt tuleb päringu koostamisel

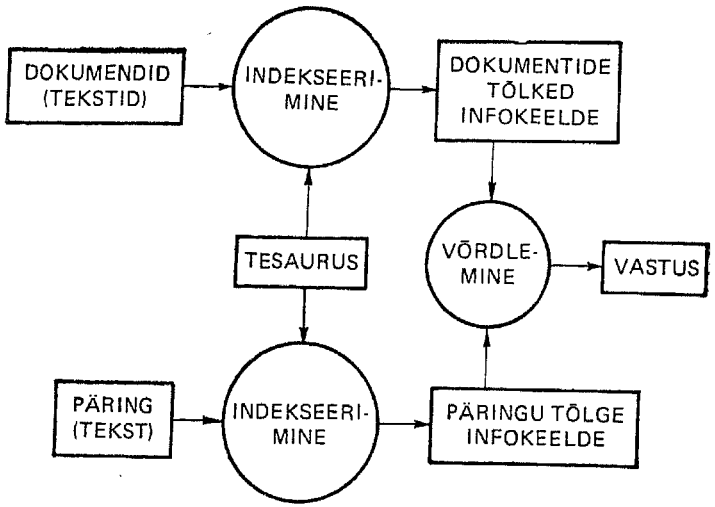
kindlaks määrata, missugusel määral kasutada tesauruses fikseeritud deskriptoritevahelisi seoseid: missuguste päringus sisalduvate mõistete puhul arvestada ka nende soomõisteid, liigimõisteid või assotsiatiivselt seotud mõisteid. On võrdlemisi ilmne, et see sõltub väga suurel määral päringu otstarbest. Kui süsteemi poole pöördub teadlane, kes vajab materjali teadusliku monograafia kirjutamiseks, siis pakub talle ilmselt huvi ka kirjandus, mis ainult kaudselt on seotud esitatud küsimusega, ja seetõttu on mõttekas kasutada lisaks soo- ja liigideskriptoritele ka assotsiatiivselt seotud deskriptoreid. Kui aga süsteemi poole pöördub spetsialist, kes vajab andmeid mingil väga konkreetsel eesmärgil (näiteks kirurg, kes valmistub teatud operatsiooniks), siis pole mingit mõtet kasutada assotsiatiivseoseid ega isegi soodeskriptoreid.

Lisaks sellele võib päringus esitada ka alternatiivseid võimalusi. Päringud ei pruugi olla ainuüksi kujul «Mind huvitavad tekstid, milles esinevad mõisted A_1 ja A_2 ja A_3 ja . . .», vaid ka näiteks: «. . . tekstid, milles esinevad mõisted A_1 ja (A_2 või A_3) ja (A_4 või A_5) . . .». Seljuhul kujutab päringu tõlge infokeelde endast loogikaavaldist, kus deskriptorid on omavahel seotud konjunktsiooni (loogika *ja*-seose) ja disjunktsiooni (loogika *või*-seose) abil. Lisaks sellele võib iga avaldises esineva deskriptori juures osutada, missuguseid temaga selle või teise seose kaudu seotud deskriptoreid tuleb otsimisprotsessis samuti arvestada.

Seda päringu tõlkimisel saadud avaldist, mida nimetatakse *otsieeskirjaks,* võrdlebki süsteem mälus leiduvate tekstide otsitunnustikega, mis mäletatavasti kujutavad deskriptoriloendeid. Sobivaks vastuseks loetakse tekstid, mille otsitunnustikus sisalduvad kõik deskriptorid, mis otsieeskirjas on seotud konjunktsiooniseose «ja» abil, ja vähemalt üks neist deskriptoreist, mis on omavahel seotud disjunktsiooniseose «või» abil. Vastusena väljastab süsteem tavaliselt viited neile dokumentidele, näiteks nende bibliograafilised kirjed, aga on ka süsteeme, mis vajaduse korral väljastavad dokumentide endi koopiad.

Infootsisüsteemi talitlust illustreerib joon. 3.

On vist pikemalt seletamatagi arusaadav, et kirjeldatud printsiipidel põhinev otsisüsteem ei saa töötada sajaprotsendilise garantiiga, see on tagada, et iga päringu puhul leitakse infomassiivist üles kõik vajalikud dokumendid ega anta välja ühtki ülearust dokumenti. Infootsisüsteemi töö kvaliteedi hindamiskriteeriumidena on levinud kaks peamist näitajat — otsingu ammendavus ja täpsus. *Ammendavuse* all mõistetakse päringule vastavate dokumentide hulka, mille süsteem mingi päringu puhul väljastab, võrreldunа tegelikult antud küsimuse kohta süsteemi infomassiivis leiduvate dokumentide hulgaga, mis inimeksperdi hinnangu järgi oleks tulnud päringu peale väljastada. *Täpsuse* all aga mõistetakse õigete, vastuseks sobivate dokumentide osakaalu kõigi



Joon. 3

dokumentide seas, mis süsteem mingi päringu peale väljastab. Mõlemad on keskmised statistilised näitajad ja nad esitatakse harilikult protsentides.

Nagu kogemused näitavad, on tänapäeva süsteemides otsingute ammendavus keskmiselt 40—50% ja täpsus 60—70%. See tähendab, et keskmine infootsisüsteem on tänapäeval oma infomassiivist suuteline üles leidma keskeltläbi poole asjakohastest dokumentidest, kusjuures väljastatavate tekstide hulgas on 30—40% ülearuseid. Need näitajad pole just eriti kõrged ning osutavad selgesti, kuivõrd praegusi infootsisüsteeme tuleb võtta üksnes kui abivahendeid infokriisi leevendamisel. Kuid neist arvudest ei maksa end ka eksitada lasta. Kui infomassiivid sisaldavad sadu tuhandeid või isegi miljoneid dokumente, siis võivad otsimistulemused tavalise käsitsimeetodi (kataloogides sorimise) puhul olla mitu korda halvemad, rääkimata tohutust erinevusest otsingule kulutatud ajas.

Senine jutt kulges võrdlemisi abstraktsel tasemel. Automatiseeritud infootsisüsteemide kasutamise tegelikud vormid ja nende roll tänapäeva ühiskonnas jäävad aga kindlasti liiga ebamääraseks, kui me eespool kirjeldatuga piirduksimegi.

Kui palju on maailmas rajatud infootsisüsteeme, selle kohta puudub ülevaade. Ainuüksi USA-s 1974. a. ilmunud «Infosüsteemide ja -talituste entsüklopeedias» on

kirjeldatud rohkem kui tuhandet automatiseeritud süsteemi. Sellest ajast on nende süsteemide loomine üha intensiivistunud. Illustreerime eelnenud kirjeldust vaid andmetega mõne praegu töötava automatiseeritud infosüsteemi kohta, ühtlasi heidame pilgu nende kasutamise organisatsioonilisele küljele.

Tänapäeval loodav automatiseeritud infosüsteem ei ole reeglina iseseisvalt funktsioneeriv üksus, vaid selle baasil luuakse terviklik infokeskus, mis lisaks infootsingule konkreetsete päringute järgi täidab mitmeid muid ülesandeid. Ja et eelnenud jutust ei jääks muljet, nagu oleksid infootsisüsteemide kasutajateks ainult «puhta» teadusega tegelevad teadlased, siis kirjeldamegi järgnevalt üht automatiseeritud infosüsteemi, mis on Nõukogude Liidus loodud terve tööstusharu, nimelt raskemasinaehituse tarvis. See harukondlik automatiseeritud teadus- ja tehnikainformatsiooni süsteem, venekeelse nimetusega *АСНТИТЯЖМАШ*[21] on mõeldud teenindama kogu kõnealuse tööstusharu ulatuses nii kollektiiv- kui individuaalabonente, ühtlasi on süsteem tööstusharu (ministeeriumi) automatiseeritud juhtimissüsteemi funktsionaalseks komponendiks. Süsteemi põhilised ülesanded on:

— tööstusharu raames ühtse infofondi komplekteerimine dokumentidega automatiseeritud töötluseks sobival kujul;

— spetsialistide ja kollektiivabonentide varustamine informatsiooniga süsteemi kompetentsi kuuluva temaatika ulatuses;

— referaat- ja bibliograafiamaterjalide väljaandmine;

— infovahetus teiste infokeskustega tööstusharu raames;

— teaduse ja tehnika arengutendentse käsitlevate infomaterjalide ettevalmistamine tööstusharu puudutavatel aladel.

Kirjeldatav infosüsteem kujutab endast tegelikult kolmetasemelist organisatsioonilist kompleksi, mis moodustub harukondlikust teadus- ja tehnikainformatsiooni keskusest, spetsialiseeritud infokeskustest, mis tegelevad tööstusharu raames peamiste teadus- ja tehnikaalade informatsiooni kogumise ja eeltöötlusega, ning ettevõtete ja organisatsioonide infotalitustest. Süsteemil on keskne infofond, mille eest kannab hoolt eelmainitud harukondlik infokeskus. Spetsialiseeritud infokeskused tegelevad vastavatele erialadele kuuluva informatsiooni kogumisega ja selle esialgse töötlemisega — bibliografeerimise, annoteerimise ja refereerimisega, ühtlasi on nad oma erialal vahendajateks abonentide ja informatsiooni keskfondi vahel. Fondi kuuluvad väga mitmesugused dokumendid: lisaks kodu- ja välismaistele teadus- ja tehnikaartiklitele ka teadusuuringute aruanded, samuti andmed käimasolevate uuringute kohta, konverentside ja sümpoosiumide materjalid, andmed leiutiste, patentide ja autoritunnistuste kohta, tehniline normdokumentatsioon

[21] АСНТИ = автоматизированная система научно-технической информации.

jne., kokku üle 40 miljoni säilitusüksuse-dokumendi.

Süsteemi abonentide — ministeeriumi, juhtivate organite, ettevõtete, üksikspetsialistide — teenindamine toimub põhiliselt kahes erinevas režiimis. Valikteadistus on abonentide pidev varustamine vajaliku informatsiooniga. Igalt abonendilt on süsteemile esitatud alaline tellimus ehk püsipäring, see on teemade ja probleemide loetelu, mille kohta abonent informatsiooni soovib. Iga infofondi tulevat dokumenti hinnatakse selle vastavuse seisukohalt püsipäringule ja seejärel informeeritakse temast abonente, kelle tellimusi dokument rahuldab. Päringurežiim on aga abonentide ühekordsete tellimuste rahuldamine.

Ka andmete väljastus on võimalik mitmel erineval viisil bibliograafiakirjete kujul (s. o. andmetena dokumendi autori(te), pealkirja ning ilmumiskoha ja -aja kohta), mikrokoopiatena, aga ka dokumentide normaalkoopiate näol.

Süsteem teenindab rohkem kui 100 000 alalist tellimust (nende seas võib muidugi olla mitu päringut ühelt abonendilt), ühekordseid päringuid aga teenindatakse iga kuu üle 2500.

Süsteemi tesauruses on fikseeritud ekvivalentsusseosed, soo- ja liigiseosed ning assotsiatiivseosed. Deskriptorite maksimaalne planeeritud arv on 150 000.

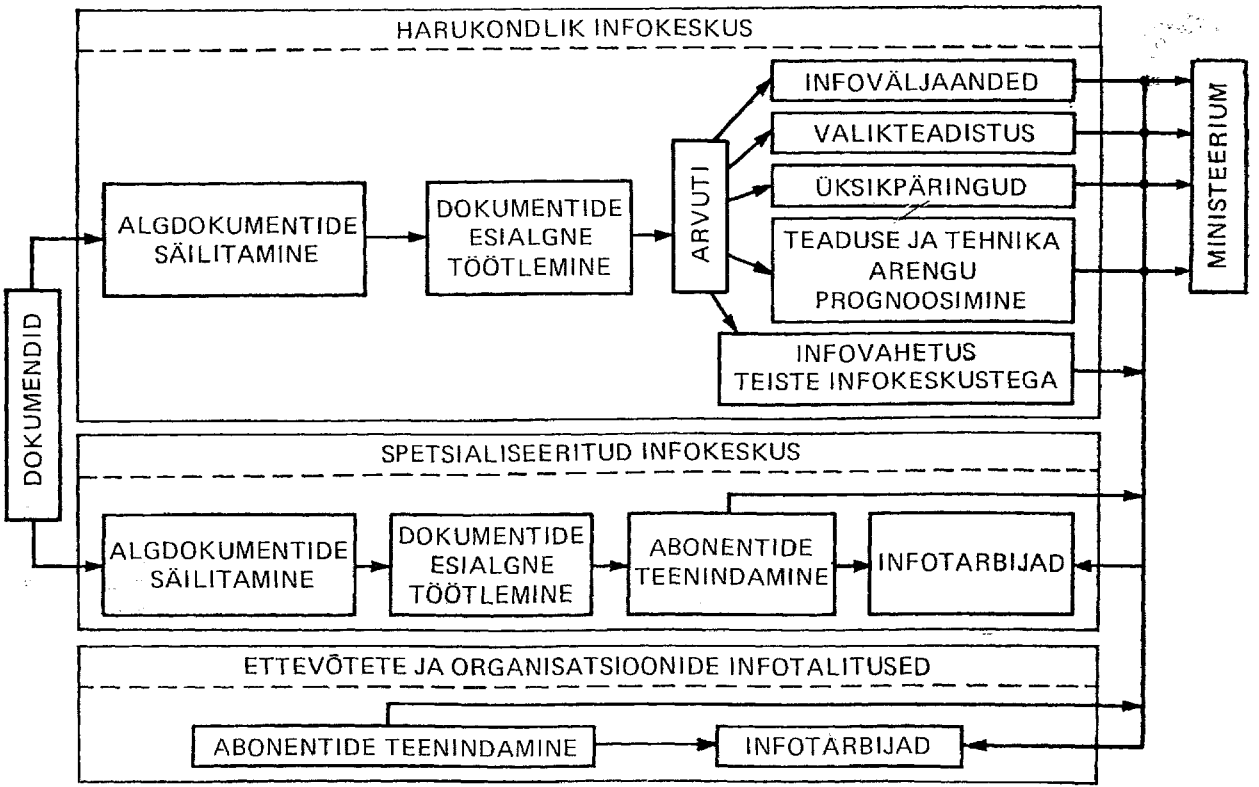
Ülevaatlikkuse huvides kujutame kirjeldatud süsteemi joonisel 4.

See süsteem pole mingi erandlik supersüsteem; kirjeldasime teda hoopis seetõttu, et ta on mitmes mõttes tüüpiline ning peegeldab tänapäeva infootsisüsteemidele iseloomulikke jooni. Üks neist seisneb selles, et loodud ja loodavad infosüsteemid ühendatakse omavahel, luues niimoodi terve infokeskuste võrgu. Igal süsteemil on oma infofond, võib olla oma infokeel, omad otsingu- ja töötlusprotseduurid. Infokeskusi ühendavate sidekanalite kaudu on igas infosüsteemis leiduv informatsioon kättesaadav ka teistele.

Nõukogude Liidus toimub infokeskuste võrgu väljaarendamine koordineeritult Riikliku Automatiseeritud Teadus- ja Tehnikainformatsioonisüsteemi (lühendatult RATTIS) raames. See on mitmetasemeline süsteem, mille allsüsteemideks on ühelt poolt liiduvabariikide vastavad automatiseeritud infosüsteemid (VATTIS-ed), teiselt poolt üleliidulised harukondlikud süsteemid, mille üht näidet kirjeldasime eespool.

Eesti Vabariiklik Automatiseeritud Teadus- ja Tehnikainformatsiooni Süsteem ehk Eesti VATTIS anti ekspluatatsiooni 1980. a. oktoobris. Süsteemi projekteerija oli Eesti Informatsiooni Instituut, ent tema talitlus hõlmab paljusid Eestis tegutsevaid infoasutusi ja raamatukogusid.

Eesti VATTIS-e põhifunktsioonid on seotud arvutil toimuva infootsinguga. Arvutis on suurte bibliograafiliste ja faktograafiliste infomassiivide moodustamine väga töömahukas ning seda pole otstarbekas teha igas vabariiklikus süsteemis eraldi, vaid hoopis üleliidulistes infokeskustes, millest igaüks


HARUKONDLIK INFOKESKUS
ALGDOKUMENTIDE SÄILITAMINE
DOKUMENTIDE ESIALGNE TÖÖTLEMINE
ARVUTI
INFOVÄLJAANDED
VALIKTEADISTUS
ÜKSIKPÄRINGUD
TEADUSE JA TEHNIKA ARENGU PROGNOOSIMINE
INFOVAHETUS TEISTE INFOKESKUSTEGA
MINISTEERIUM
DOKUMENDID
SPETSIALISEERITUD INFOKESKUS
ALGDOKUMENTIDE SÄILITAMINE
DOKUMENTIDE ESIALGNE TÖÖTLEMINE
ABONENTIDE TEENINDAMINE
INFOTARBIJAD
ETTEVÕTETE JA ORGANISATSIOONIDE INFOTALITUSED
ABONENTIDE TEENINDAMINE
INFOTARBIJAD

Joon. 4

kannab magnetlintidele teatud liiki andmeid. Eesti Informatsiooni Instituut saab üleliidulistest keskustest mitut liiki infomassiive, mida nimetatakse andmebaasideks. Üleliiduline Teadus- ja Tehnikainformatsiooni Instituut väljastab kord kuus magnetlinte, millele on kantud sama informatsioon kogu maailmas ilmunud teadus- ja tehnikapublikatsioonidest nagu selle instituudi poolt välja antavates referaatajakirjades (kaasa arvatud kogu referaaditekst). Algul andis instituut automaatika ja raadioelektroonika ning informaatika andmebaase, ent temaatika laieneb pidevalt ning hõlmab lähematel aastatel täppis- ja loodusteaduste ning tehnika kõik harud.

Üleliiduline Teadus- ja Tehnikainformatsiooni Keskus, mille ülesanne on teadusuuringute ja katsekonstrueerimistööde aruannete ning väitekirjade säilitus kui ka nende kohta teabe levitamine, väljastab samuti magnetlinte vastavate andmetega. Mahult suurim on aga patenditeabe andmebaas. Viinis asuv rahvusvaheline organisatsioon INPADOC, mille liige on ka NSV Liit, kannab magnetlintidele teavet 48 riigis väljastatavate patentide ja autoritunnistuste kohta — ilma referaatideta, kuid varustatult patendiliigituse indeksi, patenditaotluse kuupäeva, taotleja nime ning muude patendispetsialistile vajalike andmetega. Lindid saabuvad Moskvasse teadustoottmiskoondisse «Poisk», kus täiendatakse NSV Liidu ja sotsialismimaade kirjeid ning lindid paljundatakse levitamiseks. Aastas saab Eesti Informatsiooni Instituut koondiselt ligi 900 000 kirjet. Loetletud andmebaasidele tuleb edaspidi lisa.

Eesti VATTIS kasutab saabuvaid andmebaase eelkõige valikteadistuseks — infotarbijate korrapäraseks varustamiseks uue teabega vastavalt nende nõudlusele. 1982. a. sügisel oli süsteemil umbes 80 individuaal- ja kollektiivabonenti ligi 400 püsipäringuga. Kasutatav infootsisüsteem võimaldab moodustada keeruka struktuuriga päringuid, milles on näidatud huvipakkuva teabe liigitusindeksid (UDK indeks, rahvusvahelise patendiliigituse indeks jm.), erilist huvi pakkuvate autorite nimed, ajakirjanimetused, publitseerimise ajavahemik jms., eelkõige aga terminid, mis võivad esineda dokumendi pealkirjas või referaadis. Kõik otsitunnused seotakse omavahel loogikaliste kombinatsioonidena, näiteks võib ette anda terminid, mis p e a v a d kirjes esinema, teised, mis v õ i v a d seal esineda (näiteks terminisünonüümid) ning lõpuks need, mis e i t o h i seal esineda.

Üleliidulistest infokeskustest saabunud magnetlindid töödeldakse viivitamatult Riikliku Plaanikomitee Arvutuskeskuse arvutil EC—1033. Arvuti võrdleb dokumendikirjeid ükshaaval kõigi mällusalvestatud päringutega ning väljastab paberile trükitult need kirjed, mis pakuvad huvi igale üksikule tarbijale. Et kirjete ja päringute arv on väga suur, ulatub arvuti töömaht sadade tundideni aastas.

Valikteadistuse kõrval on teine

tähtis infoteenindusvorm retrospektiivne otsing andmepankades, mis sisaldavad pikema aja vältel talletatud teavet. Et üleliidulised infokeskused on väljastanud magnetlint-andmebaase alles lühikest aega, pole Eesti VATTIS-es veel jõutud rajada retro-andmepanku, kuid eeltööd selleks on tehtud. Süsteemi käsutuses oleva arvuti mälumaht pole keskse üldandmepanga jaoks piisav, seepärast moodustatakse mitu temaatilist andmepanka teadus- ja projekteerimisasutuste arvutuskeskustes. Retrootsing võib toimuda pakkrežiimis: mitme infovajaja päringud sisestatakse arvutisse üheskoos ning otsingutulemused väljastatakse trükitult. Tarbija jaoks mugavam ja tulemusrikkam on dialoogotsing, mille puhul ta suhtleb arvutiga näiteks ekraanpuldi vahendusel sidekanali kaudu. Dialoogotsingut on katsetatud ka Eesti VATTIS-es, selle rakendusväljavaated sõltuvad peamiselt sidekanalite kättesaadavusest.

Nagu teised teadus- ja tehnikainformatsiooni suursüsteemid, nii on ka Eesti VATTIS kavandatud integraalsena — temas kasutatakse ühtseid andmefonde mitmeks otstarbeks. Alates 1974. aastast koostatakse arvuti abil Eesti raamatukogudesse igas kvartalis saabunud võõrkeelsete raamatute koondnimestikke. Selle töö käigus laekunud andmestik moodustab Eesti raamatukogude fondi teatud osa koond-masinkataloogi, milles on võimalik korraldada otsingut, samuti siduda ta muude andmepankade otsisüsteemiga. On kavas hakata arvutil koostatavaid koondkatalooge välja andma mikrofišidel.

Nõukogude Liidu Riiklik Automatiseeritud Teadus- ja Tehnikainformatsiooni Süsteem on seotud Vastastikuse Majandusabi Nõukogusse (VMN) kuuluvate maade ühiste jõududega loodava Rahvusvahelise Teadus- ja Tehnikainformatsiooni Süsteemiga. Arendustööd koordineerib 1969. aastal moodustatud Rahvusvaheline Teadus- ja Tehnikainformatsiooni Keskus Moskvas. 1978. a. olid süsteemis korraldatud mitmed infoteeninduse allsüsteemid[22]:

— VMN-i liikmesmaade perioodikaväljaannete ühtne registreerimissüsteem;

— teaduslike uurimistööde spetsialiseeritud allsüsteem, mille raames vahetatakse andmeid lõpetatud ja käimasolevate teadusuuringute, dissertatsioonide jms. kohta;

— samasugune allsüsteem teadus- ja tehnikaalaste publikatsioonide kohta, mis hõlmab andmeid kõigis VMN-i liikmesmaades ilmunud publikatsioonide (raamatute, ajakirjade jne.) kohta kõigist teadus- ja tehnikaharudest.

Lõpuks tuleb öelda, et juba 1967. aastast tegutseb UNESCO raames ka ülemaailmne koostööprogramm UNISIST (täieliku nimetusega *Intergovernmental Programme for*

[22] Фенина Н. А., Калинина И. М. Развитие международной информационной системы и Международной системы научной и технической информации. — Вопросы информационной теории и практики, 1978, № 37, с. 25—37.

Co-operation in the Field of Scientific and Technological Information), mille ülesandeks on korraldada infovahetust eri maade vahel. Väljaarendatuim ja tuntuim selle programmi rakendus on perioodiliste väljaannete süsteem ISDS *(International Serials Data System)*, mis hõlmab rahvusvahelisi ja rahvuslikke infokeskusi ning mille ülesandeks on perioodikaväljaannete automatiseeritud andmefondi loomine ja hooldamine kõigil teadmisaladel (s. o. mitte ainult teaduse ja tehnika alal). Muu hulgas saab iga ISDS-is registreeritud perioodikaväljaanne standardse registreerimisnumbri koodiga ISSN, mida võime näha ka meil ilmuvate ajakirjade kaantel.

Globaalsete, tervete riikide infotalitusi haaravate süsteemide kõrval tegutseb edukalt terve rida rahvusvahelisi infootsisüsteeme kitsamatel erialadel. Niisuguse süsteemi ilmekas näide on Prantsusmaal väljatöötatud TITUS, mille infomassiivi moodustavad tekstiilialased prantsus-, saksa-, inglis- ja hispaaniakeelsed tekstid. Süsteemi ekspluateerimises osalevad Prantsusmaa kõrval Saksa FV, Belgia, Hispaania, Itaalia, USA ja Argentiina. Igal aastal kasvab süsteemi infofond 25 000 dokumendi võrra. 1972. aastast kuulub süsteemi koosseisu referaatide automaattõlke allsüsteem, mistõttu päringu võib esitada ükskõik missuguses loetletud neljast keelest ning vastuseks saadakse soovi korral dokumentide referaadid samas keeles. Umbes kümnelauselise referaadi tõlkimiseks — igas lauses keskmiselt 20 sõna — kulub kõigest 2—3 sekundit.[23] Tõlgete kvaliteet on küll kaugel täiuslikkusest, kuid oma otstarbe nad kahtlemata täidavad. Süsteemi keelevarasse kuulub erilisel viisil struktureeritud tesaurus, neli automaatgrammatikat (üks iga eespool loetletud keele jaoks) ning neli tõlkegrammatikat masintõlke allsüsteemi jaoks.

Infootsisüsteemide süvenev integratsioon nii rahvuslikes kui rahvusvahelistes mastaapides on muutnud äärmiselt aktuaalseks tõhusate infokeskustevaheliste sidesüsteemide rajamise küsimuse. Süsteemide endi operatiivsusest on vähe kasu, kui infovahetus nende vahel peaks kulgema näiteks tavalise postside kaudu. Kuivõrd tõsiselt sellesse küsimusse suhtutakse, näitab Euroopa Majandusühenduse maades loodud operatiivne infoteenindus-sidesüsteem Euronet DIANE, mille projekteerimiseks ja väljaehitamiseks on kulutatud ligikaudu 20 miljonit dollarit. Euronet DIANE ülesanne on tagada Euroopa Majandusühenduse maades terminalide (kuvarite või teletaipide) vahendusel kõigile infovajajatele vahetu juurdepääs mittesalastatud teadus-, tehnika- ja majandusandmetele mitme maa infokeskustes. Võrk avati avalikuks kasutamiseks 1979. a. novembris, 1982. a. ühendas ta enam kui 40 elektronarvutit — andmebaaside hoidjat — umbes 2500 terminaliga, kommutatsioonikeskused on Lon-

[23] Systèmes TITUS II et III. Institut Textile de France. 1976.

donis, Pariisis, Frankfurdis Maini ääres ja Roomas. 1982. aastal korraldasid Euronet DIANE abonendid keskmiselt 10 000 dialoogotsingu seanssi iga kuu.[24]

2.2. Automatiseeritud infootsisüsteemide arengutendentsid

Automatiseeritud infootsisüsteemid on keele automaattöötluse süsteemidest esimestena projektide ja katsetuste järgust jõudnud massilise tööstuseksploatatsioonini. Ehkki otsisüsteemi eelmises jaotises kirjeldatud tüüpmudel on keelevara ja «keeleliste võimete» poolest võrdlemisi primitiivsel tasemel, on sellistegi süsteemide kasutamisest infoprobleemide lahendamisel tohutult abi. Ega siis ilmaasjata kulutata nende väljatöötamisele miljoneid dollareid ja rublasid.

Samal ajal otsitakse teid otsisüsteemide keeleliste võimete arendamiseks, täiustades nendes kasutatavaid tekstitöötluse meetodeid. Mõningaid neist meetoditest vaatleme käesolevas jaotises. Me ei pea silmas teoreetilisi uuringuid, mille eesmärgiks on ehtsas mõttes teksti mõistmise õpetamine arvutile — sellele teemale on pühendatud omaette peatükk —, vaid keele automaattöötluse vorme, mida võib lugeda praegu infootsisüsteemides kasutatavate meetodite otseseks edasiarenduseks ning mille väljatöötamisel orienteerutakse nende kiirele ja massilisele rakendatavusele infoteeninduse praktikas. Vaatleme lähemalt kaht olulisemat suunda: automaatrefereerimist ning faktiotsisüsteemides kasutatavaid tekstitöötlusmeetodeid.

Automaatrefereerimise (ka *raalrefereerimise* ehk *masinrefereerimise)* all mõistetakse mingi teksti, näiteks teadusliku artikli referaadi koostamist elektronarvuti abil. Referaat on teksti lühike sisukokkuvõte, mille põhjal on võimalik otsustada lähtetekstis käsitletud probleemide, esitatud seisukohtade ja tulemuste üle. Teadus- ja tehnikakirjanduse refereerimine on töömahukas ja kulukas, ent selleta pole mõeldav tänapäeva infoteenindus. Nõukogude Liidus annab ainuüksi Üleliiduline Teadus- ja Tehnikainformatsiooni Instituut välja 48 loodus- ja rakendusteaduste referaatajakirja, millest igaüht ilmub 12 numbrit aastas. Uhiskonnateaduste ning meditsiini alal on omad referaatväljaanded, mis kõik kuuluvad enim kasutatavate väljaannete hulka. Seetõttu on arusaadav, kuivõrd suurt tööjõu-

[24] Täielikumaid andmeid Euronet DIANE struktuuri, tehniliste näitajate ja andmebaaside kohta võib leida artikleist H. Унгерер, EURONET: оперативная система информационного обслуживания стран — членов ЕЭС: возможности и перспективы. — Бюллетень ЮНЕСКО для библиотек, 1977, № 3, с. 116—138;
Ф. Т. Ф. Келли. Европейский унифицированный проект — EURONET DIANE. — Проблемы МСНТИ/МЦНТИ. М., 1981, № 2, с. 165—172.

ja ajasäästu see annaks, kui referaatide koostamise saaks jätta arvuti hooleks.

Esimesed automaatrefereerimismeetodid pakuti välja juba 1950-ndate aastate lõpus. Klassikaliseks kujunes USA teadlase H. P. Luhni poolt pakutud meetod, mille kohaselt teksti lugev elektronarvuti peab sellest välja noppima laused, mis esitavad artikli sisu seisukohalt olulisi andmeid. Oluliste lausete äratundmiseks saab kasutada statistilisi meetodeid. Luhn lähtus kahest eeldusest: esiteks, sõnad (terminid), mis esinevad tekstis kõige sagedamini, esitavad ühtlasi sisuliselt kõige olulisemaid mõisteid; teiseks, mida rohkem selliseid kaalukaid termineid sisaldab mingi lause, seda olulisem on ka lause tervikuna. Referaadi koostamine taandus niisiis lausete väljavalimisele, mis sisaldavad piisaval arvul sagedasi sõnu.

Siit pärinebki tänapäeval kasutatavate referaadi automaatse koostamise meetodite põhimõte — kindlate lausete väljavalimine refereeritavast tekstist enesest, lähtudes lausete valikul temas esinevaist sõnadest. Kuid sõnade ja lausete olulisuse määra hindamisel on puht statistilised meetodid asendunud märksa sisulisematega. Hinnatakse sõnade tähendust, nende positsiooni ja omavahelisi seoseid lausetes. Võetakse arvesse ka lausete paiknemist tekstis — näiteks asjaolu, et refereerimise seisukohalt huvipakkuvaimad andmed sisalduvad enamasti artikli alguses (kus esitatakse uurimuse eesmärgid) ja lõpus (kus esitatakse tulemused), niisamuti struktuursete alajaotuste alguses ja lõpus.

Konkreetsuse huvides kirjeldame järgnevalt S. Karasjovi poolt loodud automaatrefereerimise süsteemi ANASIN (АНАСИН), mida võib lugeda heas mõttes tüüpiliseks tänapäeva taseme esindajaks. Selles on referaadi koostamine jagatud kaheks etapiks: esiteks lähteteksti analüüs, mille käigus selgitatakse välja elemendid, mis peavad kuuluma referaati, ja teiseks referaaditeksti sünteesimine neist elementidest.

Analüüsialgoritmi ülesandeks on jagada tekstis sisalduvad laused kõigepealt kolme rühma, et neid siis erineval viisil edasi töödelda: need on laused, mis tingimata peavad kuuluma referaati, laused, mis võidakse sinna lülitada teatud täiendavate tingimuste täidetuse korral, ning laused, mis igal juhul jäävad referaadist välja. Jaotus tehakse artikli võrdlemisi üldiste struktuuritunnuste põhjal. Esimesse rühma satuvad näiteks artikli resümee ja järeldusi esitav osa (nagu näitavad uurimused, pöördub artiklit refereeriv inimene esimeses järjekorras nende osade poole). Kolmandasse rühma kuuluvad laused, mis viitavad teistele töödele või kirjeldavad teiste autorite seisukohti, samuti kuuluvad selle materjali hulka tabelid, joonised ning skeemid. Ülejäänud laused kuuluvad detailsemale töötlemisele.

Niihästi kirjeldatud liigitusprotsessi kui ka edasise analüüsi aluseks on spetsiaalsõnastik, mis on koostatud selle eesmärgiga, et te-

masse kuuluvad väljendid aitaksid lauseid vajalikul viisil klassifitseerida. Iga väljendit iseloomustatakse sõnastikus kolme liiki tunnustega — semantiliste, morfoloogiliste ja süntaktilistega. Just semantilised tunnused peavad aitama ära tunda esimese ja kolmanda rühma lauseid. Nii kuulub esimesse rühma iga lause, milles esineb konstruktsioon kategooriatest NIMI+ +SEISUKOHT: *Ivanov väidab / / arvab / usub / on seisukohal...*, *Ivanovi järgi / arvates...* jne. Morfoloogilised tunnused osutavad eelkõige sellele, missugusesse sõnaliiki kuulub sõna, kuna süntaktiliste tunnuste ülesandeks on tuua esile sõnad, mis näitavad antud lause kokkukuuluvust eelneva või järgneva lausega. Kui lause algab näiteks sõnadega *seega, järelikult, siiski, ometi, pealegi* vms., siis on selge, et ta moodustab eelneva lausega ühtse terviku.

Sünteesialgoritmi ülesanne on moodustada analüüsialgoritmi poolt referaati lülitamiseks sobivaks peetud lausetest rühmad, mis seejuures rahuldavad kindlaid semantilisi tingimusi, ning mitmesuguste teisenduste abil vormistada referaadi tekst. Teisenduste hulka kuuluvad lihtsad ümberkorraldused nagu üksikväljendite asendamine *(meie arvates → autori(te) arvates)*, aga ka tervete tekstifragmentide asendamine trafaretsete fraasidega. Näiteks püütakse eksperimentaaluuringuid kirjeldavates tekstides välja eraldada tekstiosa, milles on kirjeldatud uurimise metoodikat, ja asendatakse see tervikuna fraasiga *uurimismetoodika kirjeldus.*

Lõpptulemuseks on seega lausejärjend, mis on valitud lähtetekstist nii, et lausetes kajastuvad põhilised tekstis esitatud seisukohad, eesmärgid ja tulemused. Võimalust mööda on püütud muuta lausejärjend ka vormiliselt sidusaks, loetavaks tekstiks. Esitame ühe viieleheküljelisest artiklist saadud kolmelauselise referaadi tõlke:[25]

> Katseid tehti talinisu M 2453 kuuekümne- kuni üheksakümnepäevaste taimedega, mis olid kasvatatud sügisesel katselapil, pärast kahekümnepäevast täiendavat karastamist vegetatsioonikambris 2-kraadise temperatuuri juures. Katsetes kasutati võrseid pikkusega 2,0—2,5 cm. Autori poolt saadud tulemuste põhjal võib teha järelduse, et nii kuivatamise kui ka külmutamise korral kujunevad täiendavad vettsäilitavad jõud ainult tervetes võrsetes, mille pikkus on vähemalt 1,5 cm.

Faktiotsisüsteemides (neid nimetatakse ka faktograafilisteks infootsisüsteemideks ja andmeotsisüsteemideks) võib näha eespool kirjeldatud infootsisüsteemide edasiarendust. Senivaadeldud olid dokumendiotsisüsteemid, s. t. nad võimaldasid otsida ja vastusena väljastada üksnes viiteid dokumentidele, milles sisalduvad infotarbijat huvitavad andmed. See-

[25] Карасев С. А. Система автоматического реферирования: результаты машинного эксперимента. — Научно-техническая информация, сер. 2, 1978, № 8, с. 25—30.

vastu faktiotsisüsteem säilitab, otsib ja väljastab andmeid, fakte endid.

Et andmeid sisu järgi otsida, on kõigepealt vaja sobivat keelt nende esitamiseks. Dokumendiotsisüsteemides kasutatakse informatsiooni esitamiseks deskriptorkeeli. Dokumendi sisu kirjeldus sellises keeles kujutab endast deskriptorite (mõistete) loendit, mis tekstis esinevad, s. t. mille kohta tekstis m i d a g i räägitakse. Deskriptorkeele ainsad mõistetevahelised seosed on paradigmaatilised seosed, mille alusel mõisted koonduvad kindlatesse paradigmadesse (rühmadesse). Mingist mõistest lähtudes võime moodustada temaga soo- ja liigivahekorras olevate mõistete paradigma, temaga assotsiatiivsetes seostes olevate mõistete paradigma jne. Need vahekorrad ei võimalda esile tuua seda, kuidas mõisted on omavahel seotud vaadeldavas tekstis. Selleks et kirja panna tekstis kirjeldatud konkreetseid fakte, on aga vaja just selliseid süntagmaatilisi seoseid osutamaks, missugused tekstis mainitud mõisted on omavahel seotud ja kuidas nimelt. Näiteks väljendeil *lapse kõrvalehoidumine vanemale alimentide maksmisest* ja *vanema kõrvalehoidumine lapsele alimentide maksmisest* on täpselt ühesugune mõisteline koosseis: *laps, kõrvalehoidumine, vanem, aliment, maksmine.* Kuid kumbki väljend kirjeldab erinevat fakti, sest loetletud mõisted on neis erinevalt seotud. Faktiotsisüsteemide loomise keskseid probleemid ongi faktide esitamiseks sobivate infokeelte loomine ning meetodite väljatöötamine loomuliku keele tekstide tõlkimiseks neisse infokeeltesse. Need probleemid on omavahel otseselt seotud, sest jutt on jällegi süsteemidest, mida juba praegu või kõige lähemas tulevikus saaks hakata massiliselt kasutama. Seepärast ei saa kõne alla tulla kuigivõrd keerulise struktuuriga infokeeled, mis võimaldaksid suvalise teksti sisu täpselt kirja panna, sest niisuguste keelte tarvis ei suuda me praegu veel välja töötada küllalt tõhusaid tõlkealgoritme.

Faktiotsisüsteemide loomisel lähtutakse praegu kahest põhimõttest: esiteks, need töötatakse välja mingi suhteliselt kitsa ja kindlapiirilise ainestiku tarvis ja teiseks — vastavate tekstide sisu ei püüta tervikuna edasi anda, vaid neist nopitakse välja ainult teatud tüüpi andmed, mille kajastamine kuulub süsteemi ülesannetesse. On loodud faktiotsisüsteeme keemia, meditsiini, geoloogia, bioloogia jt. teadusalade kitsamate ainevaldade jaoks. Niisuguste tekstide puhul on võimalik võrdlemisi täpselt ette kindlaks määrata objektid, mille kohta tekstidest mingeid andmeid otsitakse, ja iga objektitüübi puhul ühtlasi määrata aspektid, mille järgi neid objekte võidakse iseloomustada ning nende kohta midagi küsida. Nii ongi kujunenud praegu kõige levinum faktiotsikeelte tüüp: o b j e k t — t u n n u s — k e e l e d. Keele üksused jagunevad kolme kategooriasse: objektid, tunnused ja tunnuseväärtused (esimene ja kolmas kategooria võivad vähemalt osaliselt kattuda). Mis valitakse objek-

tideks, mis tunnusteks ja väärtusteks, sõltub süsteemi otstarbest. Kui meid huvitab mingi keemiliste ühendite liik, siis on loomulik valida need objektideks, tunnuste hulka aga arvata nende huvipakkuvad omadused, muuhulgas näiteks see, missugustesse reaktsioonidesse need ained võivad astuda. Kuid kui meid huvitab esmajoones keemiliste reaktsioonide tüüp, siis on samavõrd loomulik objektideks valida reaktsioonid ja tunnusteks meid huvitavad reaktsioonide parameetrid, muuhulgas nende võimalikud lähteained ja tulemused. Fakti kirjeldus sellises keeles ongi objekt koos kindla tunnuse konkreetse väärtusega.

Kui deskriptorkeele puhul antakse tesauruses ette ainult deskriptorite võimalikud paradigmaatilised seosed, siis faktiotsikeeles püütakse ette anda ka võimalikud süntagmaatilised seosed: missugused tunnused saavad mingil objektil esineda ja missugune mõistete (või ka faktide) ring saab esineda ühe või teise tunnuse väärtustena.

Konkreetsuse huvides toome mõne näite. Esimene näide on geoloogiaalasest faktiotsisüsteemist, mis sisaldab andmeid maardlatest (maavarade leiukohtadest) teatud piirkonnas. Objektideks on maardlad, nende üheks olulisemaks tunnusklassiks paiknemine geograafilis-geoloogilises keskkonnas. Tunnuste väärtuseks on omakorda mitmesugused keskkonna objektid, mille suhtes maardla asukohta iseloomustatakse. Näiteks:

Objekt	Tunnused	Näited
maardla	asub väljaspool antud KO-d (keskkonnaobjekti)	maardla paikneb väljaspool intrusiivipiirkonda
	asub antud KO-st kindlas suunas kindlal kaugusel	Kimberley leiukoht asub Johannesburgist 300 km lõuna suunas
	asub antud KO-st allpool (ülalpool)	teemanditsoon on avastatud kriidikivimite kihi all
	jne.	

Teine näide on kaadriarvestuse andmesüsteemist. Selles on objektideks konkreetsed töötajad, tunnusteks töökoht, amet, kuupalk, sünniaasta jne. Niisugust andmestikku on kõige lihtsam kujutada tabelina (maatriksina), kus objektile vastava rea ja tunnusele vastava veeru lõikumiskohas on esitatud selle tunnuse väärtus antud konkreetsel objektil (vt. lk. 56).

Teksti analüüs niihästi uute andmete teatamisel kui ka päringu esitamisel on suunatud tekstis sisalduvate objekte, tunnuseid ja väärtusi tähistavate väljendite äratund-

Tunnus / Objekt	Töökoht	Amet	Sünniaasta	Kuupalk rbl.	...
Kuusik	süsteemprogrammeerimise laboratoorium	programmeerija	1949	165	...
Lepik	algoritmiteooria uurimisrühm	vanemteadur	1938	280	...
...	...	...	...	...	...

misele. On selge, et mida kitsama ala jaoks on süsteem määratud, seda lihtsam on selliste väljendite hulka vastava sõnastikuga ette anda. Kuid väljendite lihtsast äratundmisest jääb faktiotsisüsteemis tavaliselt väheks. On tarvis mingis ulatuses ka teksti morfoloogilist ja süntaktilist analüüsi, selleks et kindlaks teha, missugused väljendid tekstis on omavahel seotud ja kuidas nimelt. Selle ülesande lahendamist kergendab aga suuresti asjaolu, et mida erialasem on tekst, seda kitsam on ka tüüpiliste keelekonstruktsioonide hulk ja seda šabloonsemad nad on. Morfoloogiline ja süntaktiline analüüs orienteerub neile šabloonsetele konstruktsioonidele, ülejäänud mitmekesised, kuid suhteliselt harva esinevad konstruktsioonid jäetakse lihtsalt kõrvale.

Teine võimalus on esitada süsteemis teatud mudel tegelikkuse fragmendist, mis moodustab süsteemi ainevalla, ja selles osutada, missugustes seostes saavad olla tegelikkuse enda objektid. See on probleemi s e m a n t i l i n e l a h e n d u s: sõnade semantika ütleb, missugustesse seostesse on võimalik paigutada need sõnad ühes või teises lauses, ilma et me arvestaksime sõnade täpset morfoloogilist vormi või lause süntaktilist ehitust. Kui me näiteks teeme kindlaks, et lause koosneb sõnadest *hobune, vedama* ja *vanker,* siis ütleb meile juba sõnade semantika ja vastav «maailmamudel» ka ilma morfoloogilise ja süntaktilise analüüsita, milles saab seisneda lause poolt kirjeldatud fakt — nimelt et hobune veab vankrit ja mitte vastupidi. Seda viimast lähenemist on kasutatud näiteks eespool kirjeldatud kaadriarvestussüsteemis. Tänu sellele võib süsteem «mõista», s. o. õigesti analüüsida isegi niisuguseid suhteliselt keerulisi küsimusi, nagu «Kes programmeerijatest vanusega alla 35 aasta saab kuus rohkem palka kui sama vanad programmeerijad keskmiselt?» Mainitud süsteem sisaldab allsüsteemi, mis võib teha mitmesuguseid vastamisel vajalikuks osutuvaid arvutusi.[26]

[26] Kõnealust süsteemi on kirjeldatud artiklis Д. Я. Левин, А. С. Нариньяни. Экспериментальный минипроцессор: семантически ориентированный анализ. — В кн.: Взаимодействие с ЭВМ на естественном языке. Новосибирск, 1978, с. 223—233.

Et teksti automaatsest morfoloogilisest, süntaktilisest ja semantilisest analüüsist räägime põhjalikumalt järgmistes peatükkides, siis piirdume siin üldsõnalise kirjeldusega.

Lõpetuseks tuleb märkida, et ehkki faktiotsisüsteemide loomine on viimastel aastatel silmanähtavalt intensiivistunud ning muutunud teatud mõttes isegi moeasjaks, on praktikasse juurutatud väga väheid süsteeme, mis täies ulatuses orienteeruvad tekstide automaatsele analüüsile. Valdavas enamikus indekseerib faktiotsisüsteemides lähtetekste inimene, või siis on tema ülesanne teha tekstist rangelt etteantud formaadi järgi referaat, mida seejärel automaatselt analüüsitakse (indekseeritakse). Kuid eks selleski avaldu keele automaattöötluse objektiivne keerukus ja raskus.

3. Masintõlge

Ma olen sellel arvamusel, et tõlge ühest keelest teise — see on sama mis flaami vaip pahupoolelt vaadatuna: muster on küll enam-vähem näha, aga õiged värvid on kadunud.

Cervantes, «Don Quijote»

Tulles infootsisüsteemide juurest masintõlkesüsteemide juurde, jõuame hoopis teistsuguse iseloomuga uurimisalale.

Pole kahtlust, et praktiline vajadus tekste automaatselt ühest keelest teise tõlkivate süsteemide järele on suur ja kasvab sedamööda, kuidas suureneb eri keeltes esitatava teadusliku, tehnilise, majandusalase jne. informatsiooni hulk. Kuid kui infootsisüsteemide puhul on puht keeleline problemaatika võrdlemisi tagasihoidlik ning nende süsteemide konstrueerimismeetodeid võib mõnes mõttes vaadelda isegi kui võtteid keele analüüsi tegelikest probleemidest kõrvalehiilimiseks, siis masintõlkesüsteemides pole keele automaatanalüüsi ja -sünteesi probleemide lahendamise eest enam mingit pääsu. Seetõttu on käesolevas peatükis kesksel kohal keeleprobleemid — lausete analüüsi ja sünteesi meetodite olemus, samuti keeleväljendite struktuuri kirjeldamise formaalsed masinmeetodid.

Teiseks, masintõlge on «minevikuga» ala, sellel on olnud omad tõusud ja langused, hiilge- ja ebasoosinguperioodid. Kindlasti meenub mõistete *elektronarvutid* ja *keel* koosesinemisel paljudele praegugi kõigepealt masintõlge. 1950-ndatel aastatel, kui teadmised elektron-

arvutitest ja nende kasutamisvõimalustest üldse laiemalt levima hakkasid, oli just masintõlge arvutite võimaliku rakendusena väljapakutud aladest üks kõmulisemaid. Ometi leidub maailmas praegu vaid üksikuid süsteeme, mille puhul võib rääkida «tööstuslikust ekspluatatsioonist». Samal ajal pole selle alastest püüdlustest sugugi loobutud, pigem vastupidi — masintõlkespetsialistide optimism on tänapäeval märksa suurem kui näiteks kümme aastat tagasi. Seetõttu väärib omaette käsitlust ka masintõlkesüsteemide ajalugu, eriti kui silmas pidada, et selles käänulises protsessis kajastub mõneti vaadete muutumine keelele ja keele automaattöötlusele tervikuna.

3.1. Ajalugu ja probleemid

Nagu väidetakse, tekkis masintõlke idee juba 1946. aastal USA teadlaste W. Weaveri ja A. D. Boothi vahelises vestluses, kus arutati elektronarvutite rakendamise uusi võimalusi. 1949. aastal ilmus W. Weaveri artikkel «Tõlkimine», milles laiemale avalikkusele tutvustati võimalust kasutada arvuteid tekstide tõlkimiseks ühest keelest teise.[27] See idee võeti entusiasmiga vastu ning Washingtoni ja California ülikoolides kui ka Massachusettsi tehnoloogiainstituudis alustati peagi vastavaid uuringuid. 1954. aastal demonstreeriti esmakordselt masintõlkesüsteemi, mis tõlkis inglise keelest vene keelde, kasutades 250-sõnalist sõnastikku ja kõigest kuut süntaksireeglit. Samal aastal hakati välja andma masintõlke probleemidele pühendatud eriajakirja «Mechanical Translation».

Nõukogude Liidus algasid masintõlkealased uuringud 1954. aastal, mil Moskvas NSVL TA Täppismehaanika ja Arvutustehnika Instituudis alustas tööd vastav uurimisrühm D. Panovi ja I. Muhhini juhtimisel. 1955. aasta detsembris jõuti esimese eksperimentaalse masintõlkeni inglise keelest vene keelde. 1956. aastal viidi läbi prantsuse-vene masintõlkeeksperimendid NSVL TA Steklovi-nim. Matemaatika Instituudis, kus töid juhtisid A. Ljapunov ja O. Kulagina. Neist teine on tänapäeval üks meie tunnustatumaid masintõlkespetsialiste. Mõlema rühma poolt korraldatud katsete tulemused ja neist saadud kogemused avaldati 1956. aastal ühises kogumikus «Ühest keelest teise tõlkimise automatiseerimine».

Umbes selleks ajaks olid selgunud ka esimesed raskused ja hakkas arenema üha ägedamaks muutuv diskussioon, kus ühel pool seisid masintõlke pooldajad — entusiastid, teisel pool need, kes püüdsid tõestada tõlkimise automatiseerimise mittevajalikkust, praktilist teostamatust või koguni põhimõttelist võimatust.

[27] Artikli venekeelne tõlge on avaldatud kogumikus Машинный перевод. М., 1957, с. 15—32.

Tuleb silmas pidada, et kõik arvutite abil tehtavad tööd on tänapäevalgi suhteliselt kallid, seda enam ajal, kui elektronarvutite seeriatootmine oli alles alanud. Iga arvutite kasutamist eeldav uurimisprojekt oli mõeldav vaid ulatuslike riiklike või muude subsiidiumide korral. Seetõttu ei olnud vaidlus masintõlke ümber üksnes teoreetiline arutelu, vaid ühtlasi võitlus teaduslike uurimistööde jaoks eraldatavate summade jaotamise eest. See seletab mõningal määral diskussiooni ägedust, samuti ka seda, miks selle käigus kasutati palju sisult lausa absurdseid, ehkki enamvähem soliidsesse teaduslikku rüüsse riietatud argumente. Näiteks püüti masintõlke võimatust lähemas tulevikus tõestada järgmise arutluskäiguga. Paljud keeleteadlased ja filosoofid (nimetati sel ajal autoriteetideks loetud nimesid) on veenvalt tõestanud, et mõtete väljendamise ja suhtlemise vähim üksus on lause, mitte sõna. Masintõlkele rakendatult tähendab see, et arvuti mälus rakendatavateks üksusteks peavad olema mitte sõnad, vaid laused. On võimalik moodustada 10^{50} kuni 20 sõna sisaldavat lauset (veelgi pikemaid lauseid aga märksa rohkem); masintõlkesüsteem peaks seega säilitama mälus vähemalt $2 \cdot 10^{50}$ lauset pluss viited selle kohta, missugused ühe keele laused on missuguste teise keele lausete tõlkevasted. See aga tähendab, et efektiivsest masintõlkest lahutab meid vähemalt 45 suurusjärku, sest olemasolevate arvutite mäluplokkide maht on ülalmainitud arvust selle võrra väiksem.[28] See on omapärane arutlus, kus faktid omaette võetult just nagu paika peavad, kuid üleminekud ühelt faktilt teisele on täiesti põhjendamatud. Eelkõige käib see arutluskäigu avasammu kohta. On tõsi, et inimesed suhtlevad lausete ja mitte sõnade abil ning et mõttelist iseseisvust ja terviklikkust võib suhtlemises omada alles lause. Kuid siit ei saa järeldada, et laused peavad olema ka need vähimad üksused, mis tuleb mälus säilitada ja sealt valmis kujul väljastada, kui vastavat mõtet tahetakse väljendada, mõista või tõlkida. Laused moodustatakse sõnadest kindlate reeglite järgi ja neid reegleid kasutatakse nii oma mõtete väljendamisel kui ka teiste poolt moodustatud lausete mõistmisel, seega ka tõlkimisel. Nii on see inimeste puhul; väita, et neid reegleid pole põhimõtteliselt võimalik välja selgitada ja mingilgi määral formaliseerida, on sama kui väita, et nad kuuluvad mingisse müstilisse, inimtunnetusele mittealluvasse maailma. Seepärast omab ka arv 10^{50} ja kogu sellele üles ehitatud arutlus vähe mõtet — rääkimata sellest, et ka arv ise on üpris meelevaldne, sest ta esitab teoreetiliselt võimalike lausete hulka, mis ei arvesta mingeid grammatilisi ega tähenduslikke kitsendusi sõnade kombineerimisel lausetes.

Toodud arutlus spekuleeris raskustega, mille juured ei peitunud keele olemuses, vaid hoopis neis konkreetsetes meetodites, mille

[28] Таубе М. Вычислительные машины и здравый смысл. М., 1964, с. 37.

abil esimestes masintõlkesüsteemides keeltevahelistest erinevustest püüti jagu saada. Need süsteemid orienteerusid täielikult «jõumeetodile», sõnasõnalisele tõlkimisele, põhitöövahend oli kakskeelse sõnastiku masinvariant. Algul olid sõnastikku lülitatud ka sõnade kõik morfoloogilised vormid. Hiljem leiti, et on otstarbekam esitada sõnastikus ainult sõnade põhivormid ning lisada talle morfoloogilise analüüsi protseduurid, mille abil tuntakse ära tekstis esinevad konkreetsed vormid. Niisiis, kui esialgu esitati vormid *raamatus, raamatutest, raamatutega* sõnastikus igaüks omaette ning igaühe juures anti tema võimalikud vasted keeles, millesse sooviti tõlkida (nn. väljundkeel), siis hiljem oli sõnastikus vaid põhivorm *(raamat)* või mitu põhivormi, mida ei saa moodustada kindlate reeglite järgi, kuna morfoloogilise analüüsi algoritmi ülesanne oli kindlaks teha, et näiteks sõnavorm *raamatutega* on sõna *raamat* mitmuse kaasaütlev kääne. Kuid tõlkimisel püüti vastavused kehtestada ikkagi sõnavormide tasemel. Lause süntaktilist ehitust, sõnade seostumist teiste sõnadega lauses arvestati minimaalsel määral. Näiteks 1954. aastal korraldatud esimeses eksperimendis kasutati vaid kuut süntaksireeglit, sellal kui tänapäeva täiuslikemates süsteemides ulatub süntaksireeglite arv sadadesse või isegi tuhandetesse, sõltuvalt sellest, mida loetakse üheks reegliks. Lausete tähenduslikku aspekti, semantikat, ei arvestatud aga praktiliselt üldse.

On arusaadav, et niisuguse lähenemise puhul muutus sõnade mitmetähenduslikkuse probleem peagi ületamatuks. Niipea kui ühe keele sõnale vastab teises keeles mitu sõna, seistakse küsimuse ees, missugune neist vastetest sobib antud juhul. Tarvitseb vaid põgusalt lehitseda mõnd kakskeelset sõnastikku veendumaks, et olukord, kus ühele sõnale vastab teises keeles mitu erinevat sõna, on pigem reegel kui erandjuhtum. Seejuures pole enamasti sugugi ükskõik, missugust neist sõnadest kasutada lähtekeele sõna vastena. Näiteks eesti keele sõnale *pikk* vastab vene keeles kord *длинный*, kord *высокий;* pikk mees on *высокий мужчина*, mitte *длинный мужчина*. Eesti keele sõnale *suur* vastab inglise keeles kord *big*, kord *large*, kord *great*, vene keeles kord *большой*, kord *великий*.

Harvemini esinev, kuid seda teravam on homonüümide probleem — juhtumid, kus ühele sõnale vastab teises keeles mitu sõna, mis tähenduslikult pole üldse seotud. Kui eeltoodud näidete puhul on väära sõna kasutamise korral (*длинный мужчина*) õige vorm enamasti taastatav, siis homonüümide puhul võib väärvaste kasutamine tekitada täieliku mõttetuse. Vene keele sõna *ключ* tähendab nii võtit kui ka allikat. Sõltuvalt kasutatud vastest muutub täiesti lause tähendus, milles see sõna esineb.

Esitame järgnevalt tabeli kujul konkreetse näite: mis saab, kui sõnasõnalisel meetodil püüda tõlkida saksakeelne lause *Das ist ein grosses Schloss 'See on suur loss'*

Sõna-vorm	Sõnaliik	Morfoloogiline info		Ingliskeelne vaste					Eestikeelne vaste			
das	määrav artikkel	ainsus kesksugu	nominatiiv / akusatiiv	the					—			
	näitav asesõna	ainsus kesksugu	nominatiiv / akusatiiv	this (one)	that (one)	he	she	it	too see	tolle selle	toda seda	ta/tema teda
	siduv asesõna	ainsus kesksugu	nominatiiv / akusatiiv	who / whom	which	that			kes mis	kelle mille	keda mida	
ist	tegusõna	kindel kv. ainsus olevik	3. pööre	is					on			
ein	määrsõna	seotud väljend		—					—			
	tegusõna lahutatav eesliide			—					—			
	umbmäärane artikkel	ainsus meessugu	nominatiiv	a an					—			
		ainsus kesksugu	nominatiiv / akusatiiv									
	umbmäärane asesõna	—	—						üks	keegi		
	arvsõna	meessugu / kesksugu	nominatiiv / nominatiiv / akusatiiv / muutumatu	one					üks	ühe		üht

Sõnavorm	Sõnaliik	Morfoloogiline info		Ingliskeelne vaste					Eestikeelne vaste				
grosses	omadussõna	ainsus kesksugu algvõrre	nominatiiv akusatiiv	big	large	great	tall		suur	suure		suurt	
Schloss	nimisõna	ainsus	nominatiiv akusatiiv	castle	lock				loss lossi	lukk luku lukku			
		kesksugu	daativ										
	tegusõna	kindel kv. minevik ainsus	1. pööre 3. pööre	shut	close	lock	conclude		sulgesin sulges	lukustasin lukustas	järeldasin järeldas		

inglise või eesti keelde.[29] Kes saksa või inglise keelt ei valda, sellele jääb näide teatud määral segaseks, kuid peamine fakt — sõnasõnalise vastavuse puudumine keelte vahel — peaks kõigile selgeks saama. Nii võib tabelist välja lugeda, et saksa keele sõna *das* saab omaette võetult inglise keelde tõlkida vähemalt kümnel, eesti keelde koguni viieteistkümnel erineval viisil.

Kuivõrd tõsiseks probleemiks osutus üksühese vastavuse puudumine keelte vahel sõnavara ja morfoloogia tasemel, seda näitavad ilmekalt 50-ndate ja 60-ndate aastate masintõlke ühe juhtiva teoreetiku Y. Bar-Hilleli sõnad aruandest USA Rahvuslikule Teadusnõukogule 1959. aastal (ajal, kui sõnalise tõlke meetodeid oli juba mõnda aega täiendatud süntaktilise analüüsi meetoditega). Bar-Hillel ütles, et ei ole olemas ja lähemal ajal ka ei paista kusagilt mingit täiesti automaatset metoodikat, mis võimaldaks olemasolevatel (mitteõppivatel) elektronarvutitel lahendada sõna *pen* polüseemiat niisugustes lausetes nagu *The pen is in the box* ja *The box is in the pen* samasugusel viisil, nagu seda kiiresti ja eksimatult teeks inimene.

Asi on selles, et sõna *pen* tähendab inglise keeles niihästi (kirjutus)sulge kui ka karjatara (aedi-

[29] Bruderer H. Datenverarbeitung und Sprachwissenschaft. — Rmt.: Sprache und Computer. Zürich, 1977, S. 17—18. (Sealsed prantsuse keele andmed asendasime eesti keele omadega.)

kut), *box* tähendab karpi, kasti. Niisiis inimene tõlgiks esimese lause *Sulg on karbis,* teise aga *Kast on karjataras.* Morfoloogiliste ja isegi süntaktiliste kriteeriumide järgi ei ole aga midagi viga ka lausetel *Karjatara on kastis* ja *Kast on sules.* Viimased laused on aga sisuliselt nonsensid.

Intuitiivselt pole eriti raske taibata, mis on nende raskuste põhjus ja kust otsida väljapääsu olukorrast. Mitut erinevat tähendust omab sõna üksikult, omaette võetuna. Enamus sõnu kaotab inimese jaoks mitmetähenduslikkuse, kui ta esineb konkreetses lauses, mida omakorda on kasutatud konkreetses situatsioonis. Näiteid selle kohta tõime juba esimeses peatükis, sedasama tõendavad äsja vaadeldud näited. Kõige lihtsamad on juhud, kus eri sõnad ühtivad ainult mõnedes grammatilistes vormides, nagu *tuli* (nimisõna) ja *tuli* (tegusõna), niisamuti *sai* ja *sai*, *viin* ja *viin*, *ilus* (omadussõna) ja *ilus* (nimisõna seesütlevas käändes). Siin aitab harilikult raskustest üle grammatilise konteksti analüüs, tehes kindlaks, missuguste teiste sõnade vormidega antud sõna lauses vahetult koos esineb. Niisuguse analüüsi põhjal võime näiteks öelda, et lauses *Poiss tuli koju* on *tuli* tegusõnavorm, ilma et me hakkaksime süvenema sõnade konkreetsematesse iseärasustesse. Niisamuti võime eeltoodud saksakeelse lause *Das ist ein grosses Schloss* puhul vahetu ümbruse analüüsi põhjal öelda, et *das* ei saa siin olla artikkel ega siduv asesõna.

Enamasti ei saa raskustest üle niivõrd lihtsate vahenditega. Selleks et kindlaks teha, missuguste teiste sõnadega meid huvitav sõna on lauses grammatiliselt seotud, tuleb harilikult välja selgitada lause kui terviku süntaktiline struktuur, s. o. määrata sõnadevahelised seosed kogu lause ulatuses. Sageli ei aita ka see: võime lause grammatilise (süntaktilise) ehituse täpselt ja üheselt kindlaks teha, aga lause on ikkagi mitut moodi tõlgitav. Bar-Hilleli näited *The pen is in the box* ja *The box is in the pen* on nimelt sellised. Siin on vaja lause s e m a n t i l i s t analüüsi. Tuleb appi võtta sõnade tähendused ning lisaks on vaja reegleid, mis määraksid, missugused tähenduste kombinatsioonid on mõttekad ja missugused ei ole. Ainult niisuguste reeglite abil on võimalik välistada tõlkeid nagu *Kast on sules* või *Karjatara on kastis.* Peale selle tuleb silmas pidada, et lisaks juhtudele, kus lause enda põhjal on selge tema mõttekus või mõttetus, on terve müriaad olukordi, kus lause mõttekus või mõttetus on suhteline — ühes kontekstis on ta mõttekas, teises mitte.

Niisiis sai selgeks, et mitmetähenduslikkuse probleemi lahendus peitub kontekstis. Just kontekst ütleb, missugune kõigist võimalikest tõlgitsustest valida konkreetsel juhul — ütleb üksiksõnade, sõnaühendite, lõpuks ka lausete puhul. Kuid mida kontekst endast kujutab ja k u i d a s teda arvestada? Püüd vastata nendele küsimustele viis masintõlkespetsialistid keele süntaksi- ja semantikaprobleemide rägastikku. Nendes prob-

leemides orienteerumiseks puudus aga ettevalmistus.

Nii oldi paratamatult sunnitud tunnistama, et jõumeetodi rakendamine ning panus elektronarvuti suurele mälule ja kiirusele ei ole viinud loodetud sihile, vähemalt mitte nii ruttu ja sirgjooneliselt, kui algul kujutleti. Leiti, et tahestahtmata on vaja tõsisemalt süveneda keele ehitusse.

Kriis saabus 50-ndate aastate lõpul, siis kui alustati esimesi tõsisemaid katseid lülitada masintõlkeprogrammidesse lisaks morfoloogilisele analüüsile ka süntaktilise analüüsi komponent. Selgus, et automaatse süntaktilise analüüsi meetodid on võrreldamatult keerulisemad, seetõttu ka kulukamad, aeganõudvamad ja ebakindlamad kui morfoloogilise analüüsi meetodid, ja teiseks, et teadmised keele süntaktilise tasandi seaduspärasustest on selleks liiga lünklikud ja ebatäpsed, et loota nende peatsele algoritmimisele.

Kõik see oli taustaks kahele sündmusele, millest üks toimus Nõukogude Liidus, teine USA-s ja mis küllalt pikaks ajaks vajutasid pitseri masintõlke alal tehtavale tööle, eriti aga selle töö mainele avalikkuse silmis.

1961. aastal toimus Moskvas masintõlke, infotöötluse ja automaatse kõneanalüüsi üleliiduline konverents. Selle lõppistungil deklareeriti, et masintõlge kui praktiline ülesanne tuleb päevakorrast maha võtta ja kõik jõupingutused suunata tõlkimise ning üldse keeleanalüüsi teoreetiliste aluste uurimisele.

Ameerika Ühendriikides läks asi veelgi kaugemale. Moodustati soliidne riiklik komisjon masintõlke perspektiivikuse hindamiseks, mis uuris põhjalikult masintõlke olukorda USA-s ja ühe konkreetse süsteemi põhjal (selleks oli õhujõudude valitsuses töötav süsteem SYSTRAN, millest meil tuleb juttu hiljem) võrdles kulutusi masintõlke ja inimtõlke puhul. Komisjon otsustas, et tõlgete madala kvaliteedi ja kalliduse tõttu ei tasu masintõlge end ära. Vajaliku kvaliteediga tõlgeteni jõudmiseks tuleb arendada fundamentaaluuringuid lingvistika ja matemaatika piirimail.

Komisjoni aruanne ilmus trükis[30] ning loomulikult avaldas märgatavat mõju suhtumisele masintõlkesse kogu maailmas. Süsteem SYSTRAN ise elas inspekteerimise küll õnnelikult üle ja jätkas tegevust, kuid nii USA-s, Nõukogude Liidus kui ka mujal maailmas algas masintõlkealaste tööde radikaalne ahendamine. Masintõlge lakkas olemast esimese järgu tähtsusega teaduslik-tehniline ülesanne, kuigi sellega tegelesid edasi paljud uurimiskollektiivid.

70-ndate aastate alguses kasvas taas aktiivsus sel tegevusalal ja ühtlasi tõusis uuesti huvi masintõlke vastu. Peamine põhjus seisnes selles, et vahepealse kümmekonna aasta jooksul oli masintõlkealane töö ise tugevasti edasi liikunud, seda eelkõige koos vastavate

[30] Venekeelne tõlge vt. ajakirjast Научно-техническая информация, сер. 2, 1968, № 8, с. 23—36.

teoreetiliste keeleteaduslike uurimustega keele kirjeldamise täppismeetodite väljatöötamiseks. Kui 50-ndatel aastatel oli keeleteaduse keskne probleem keele fonoloogilise (häälikulise) ja morfoloogilise ehituse kirjeldamine, siis 60-ndad aastad olid peaaegu täielikult keele süntaktilist struktuuri — lausete ehitust — käsitlevate teooriate päralt, millele 70-ndatel aastatel lisandus keele semantilise struktuuri — sõnade ja lausete tähenduste — üha süvenev uurimine.[31]

*Masin*tõlkes *eneses on samuti* üle elatud mitu perioodi, ehk nagu tavatsetakse öelda — on vahetunud mitu masintõlkesüsteemide põlvkonda.

Esimese põlvkonna hulka loetakse süsteemid, milles piirduti põhiliselt morfoloogilise analüüsiga. Kohe pärast analüüsi algas lähtekeele lauselt väljundkeele lausele ülemineku ülimalt keeruline protsess. Süsteemide niisuguse ehituse tõttu oli vaja iga keeltepaari jaoks välja töötada omad analüüsi- ja sünteesialgoritmid; näiteks tõlkimaks inglise keelest vene keelde oli vaja hoopis teistsugust analüüsialgoritmi kui tõlkimiseks inglise keelest prantsuse keelde.

Teise põlvkonna süsteemides, mida hakati välja töötama 60-ndatel aastatel, toimus analüüsi- ja sünteesietappide eraldumine. Kekseks kujunes süntaktiline analüüs, mille eesmärk oli lause süntaktilise struktuuri konstrueerimine. Seejuures püüti saavutada, et süntaktilise struktuuri kirjeldus oleks võimalikult sõltumatu väljundkeelest. Sel juhul võib ühtesid ja samu analüüsialgoritme kasutada antud lähtekeelest mis tahes teise keelde tõlkimisel.

Lõpuks, kolmandasse põlvkonda arvatakse harilikult süsteemid, kus lisaks süntaktilisele analüüsile kasutatakse süstemaatiliselt semantilist analüüsi. Neid hakati ehitama alles 70-ndatel aastatel ning kogu maailmas on valminud vaid paar eksperimentaalsüsteemi, mida tinglikult võib lugeda siia kuuluvaks.

Nii masintõlke arengule kui ka tema maine paranemisele kogu maailmas on soodsalt mõjunud ka asjaolu, et masintõlge pole enam ainus ala, kus arvuteid kasutatakse keelematerjali töötlemisel. Eelmises peatükis rääkisime automatiseeritud infootsisüsteemidest ja osutasime, et nende kasutamine teaduse, tehnika ja majanduse mitmesugustel aladel on massiline. On jõutud sellistegi mitmekeelsete infootsisüsteemideni, kus kasutatakse masintõlke elemente.

Viimasel aastakümnel on hakanud kiiresti edenema põhimõtteliselt uut tüüpi süsteemide — tehisintellektisüsteemide loomine, mille hulgas keskse kohal on loomulikust keelest — tekstide sisust — aru saavad süsteemid. Nende loomisega seotud teoreetilised ja praktilised probleemid on mitmes suhtes märksa keerulisemad kui masintõlkesüsteemide puhul, ning neile ongi pühendatud meie raamatu järgmine peatükk.

[31] Läbilõike kõnealustel aastatel keeleteaduses toimunust võib leida raamatust H. Oim. Semantika. Tallinn, 1974.

Lõpuks ei saa jätta mainimata, et ka masintõlke tehniline baas — elektronarvutid — on vahepeal hiigelsammudega edasi arenenud. 50-ndate aastate esimese põlvkonna arvutite juurest on jõutud kolmanda ja isegi neljanda põlvkonna arvutiteni, mis on suutelised tegema miljoneid tehteid sekundis ja mille mälu maht (kui välismäluseadmed juurde arvata) on praktiliselt piiramatu.

Kõige selle tulemusena on 70-ndate aastate tõlkesüsteemid praktiliste võimete ja ka rentaabluse seisukohalt hoopis midagi muud kui 60-ndate aastate alguse süsteemid. Selle ilmeka näitena suutis Ameerika masintõlkespetsialist B. Dostert 1974. aastal oma süsteemiga üheainsa öö jooksul tõlkida 400-leheküljelise tehnikaalase raamatu vene keelest inglise keelde. Tõlke kvaliteet oli Dosterti enda sõnade järgi küll kohutav, kuid see hinnang puudutas peamiselt stilistilist külge. Tõlke sisuline vastavus originaalile oli sedavõrd hea, et spetsialistid võisid teda pärast kerget redigeerimist juba järgmisel päeval kasutada.

Ametliku suhtumise muutumist masintõlkesse iseloomustab kõige selgemini järgmine fakt. Aastatel 1973—1974 uuris masintõlke olukorda meie maal Riikliku Teadus- ja Tehnikakomitee erikomisjon eesmärgiga välja selgitada, kuivõrd on võimalik ja otstarbekas juba praegu asuda praktiliste masintõlkesüsteemide loomisele. Praktilise süsteemi all mõeldi säärast, mis on suuteline korrapäraselt ja massiliselt tõlkima teadus- ja tehnikaalaseid tekste, kusjuures tõlke kvaliteet võib olla suhteliselt madal: peamine on, et tõlge oleks tellijale arusaadav, vastaks sisuliselt originaalile ja oleks kasutatav informatsioonivahendina.

Komisjoni peamine järeldus oli järgmine: «Masintõlkealaste teoreetiliste ja katsetööde praeguseks saavutatud tase lubab tõstatada küsimuse masintõlke praktilisele realiseerimisele asumisest Nõukogude Liidus.»[32] Eriti märkis komisjon, et praktiliste masintõlkesüsteemide loomisele ja juurutamisele tuleb asuda kohe, ootamata probleemiga seotud teoreetiliste küsimuste kogu kompleksi lahendamist.

Komisjon formuleeris ka võrdlemisi selged tingimused, millest lähtudes hinnata masintõlke ökonoomsust:

— teksti masintöötlus olgu umbes viis korda odavam kui inimese poolt tehtav tõlge;

— teksti sisestus arvutisse olgu töömahukuselt võrreldav inimese poolt tehtud tõlke käsikirja kirjutusmasinal ümberkirjutamisega;

— tõlkimine arvutil koos teksti järelredigeerimisega võtku vähemalt kümme korda vähem aega kui sama teksti tõlkimine inimese poolt.

Enne kui asume lähemalt kirjeldama masintõlkesüsteemi ehitust, loetleme pildi täielikkuse huvides — aga ka selleks, et lugejal ei jääks muljet masintõlkest kui

[32] Котов Р. Г. Лингвистика и современное состояние машинного перевода. — Вопросы языкознания, 1976, № 5, с. 37.

varjusurmas olnud alast, mis alles viimastel aastatel on taas ellu ärganud — kõik olulisemad vahepealse perioodi jooksul Nõukogude Liidus loodud süsteemid, mis on jõudnud praktiliste katsetusteni (nimetamata süsteeme, mis praegu on alles väljatöötamisel).

1. Inglise-vene masintõlkesüsteem, loodud Moskvas NSVL TA Täppismehhaanika ja Arvutustehnika Instituudis (sellest oli juttu eespool). Esimese põlvkonna süsteem.

2. Prantsuse-vene masintõlkesüsteem, loodud Moskvas NSVL TA Rakendusmatemaatika Instituudis. Esimese põlvkonna süsteem (esimesed katsetused tehti 1956. a.).

3. Matemaatiliste tekstide tõlkimiseks loodud inglise-vene tõlkesüsteem, loodud samas instituudis eelmise süsteemi põhimõtete edasiarenduse tulemusel. Valmis 1950-ndate aastate lõpus.

4. Vene-ukraina tõlkesüsteem, mis valmis Ukraina NSV TA Küberneetika Instituudis 1960-ndate aastate alguses. Katsetati mitut varianti: ainult sõnastikule tuginevat tõlkimist, sõnastikku koos morfoloogilise analüüsiga kui ka sõnastikul, morfoloogilisel analüüsil ja süntaktilise analüüsi elementidel põhinevat tõlkimist (seega võiks öelda, et süsteem kuulus $1^1/_2$-ndasse põlvkonda).

5. Armeenia NSV TA Arvutuskeskuses väljatöötatud armeenia-vene tõlkesüsteem, mille katsetamiseni jõuti 1960-ndate aastate esimesel poolel. Sel süsteemil oli mitmeid teisele põlvkonnale iseloomulikke jooni: kasutati süntaktilist analüüsi ning analüüsi ja sünteesi etapid olid teineteisest eraldatud, kuid analüüs oli siiski orienteeritud ühele konkreetsele väljundkeelele.

6. Matemaatiliste tekstide tõlkimiseks loodud armeenia-vene tõlkesüsteem, mis sarnanes eelmisega ning töötati välja samas asutuses 1960-ndate aastate teisel poolel.

7. Patendikirjelduste tõlkimiseks mõeldud inglise-vene süsteem, mis töötati välja 1960-ndate aastate lõpul Moskvas Patendiinformatsiooni Teadusliku Uurimise Keskinstituudis.

8. Gruusia-vene tõlkesüsteem, mis loodi Gruusia NSV TA Juhtimissüsteemide Instituudis 1970-ndate aastate alguses ning vastab juba enamikule teise põlvkonna süsteemidele esitatavatele nõuetele.

9. Leningradi Riikliku Ülikooli Arvutuskeskuses väljatöötatud inglise-vene tõlkesüsteem teadus- ja tehnikaalaste tekstide tõlkimiseks. Süsteemi katsetamine algas juba 1966. aastal. Hilisemal, täiustatud kujul on talle omased kõik teise põlvkonna iseloomulikud jooned.

10. NSVL TA Rakendusmatemaatika Instituudis matemaatiliste tekstide tõlkimiseks loodud prantsuse-vene tõlkesüsteem, mis kujutab endast samas instituudis väljatöötatud eelmise süsteemi uut varianti. See 1970-ndate aastate alguses valminud teise põlvkonna süsteem on praeguseks läbi teinud põhjalikud katsetused.

Kogu maailmas loodud süsteeme ei ole mõeldav üles lugeda. Ainuüksi Ameerika Ühendriikides kes-

tavad praegu masintõlkealased uuringud enam kui kümnes kohas. Keelte arv, millest või millesse tõlkimiseks üldse on loodud masintõlkesüsteeme, on küllaltki ebatäielikel andmetel kahekümne ümber. Nende hulgas on selliseid eksootilisi keeli, nagu tai ja vietnami keel (1971.—1973. a. töötas USA-s süsteem «Logos», mis tõlkis inglise keelest vietnami keelde), samuti esperanto.

Tuntumate tänapäeval töötavate ja väljatöötamisel olevate süsteemide juurde tuleme peatüki lõpus tagasi.

3.2. Kuidas töötab masintõlkesüsteem

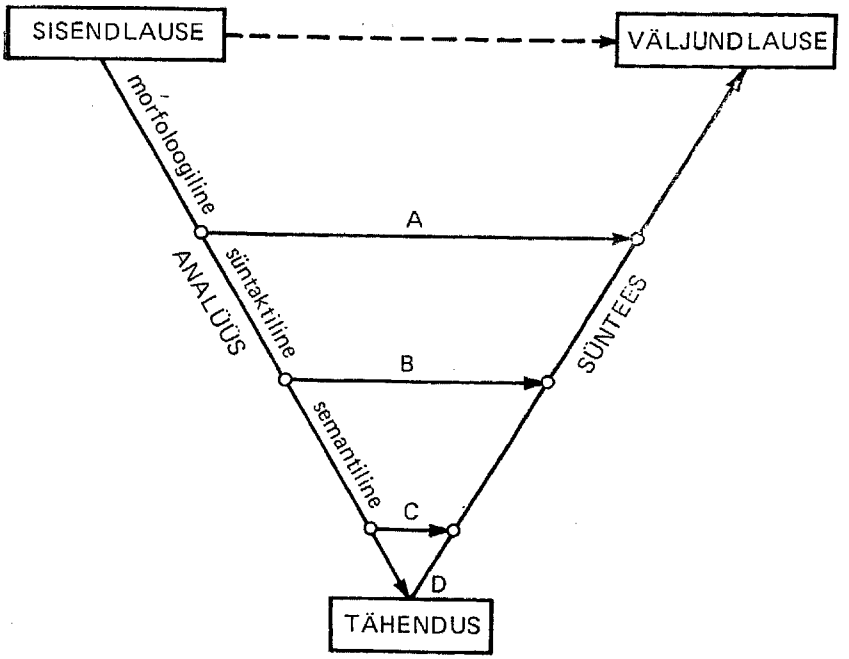
Eelnenu põhjal on ilmselt juba selge, et masintõlkesüsteemi põhiline töötsükkel jaguneb kolme etappi: sisendkeele lause analüüs, selle analüüsi tulemuseks oleva alusstruktuuri teisendamine väljundkeele vastava lause alusstruktuuriks ning antud alusstruktuurist lähtudes väljundkeele lause süntees. Üleminek ühelt keelelt teisele võib toimuda erinevatel sügavusastmetel. Seda olukorda saab illustreerida skeemiga joonisel 5.

Esimene, tähistuseta üleminek on kõige otsesem tõlge sõnavormide sõnastiku põhjal ilma igasuguse analüüsita. Üleminek A on tõlkimine ainuüksi morfoloogilise analüüsi põhjal, B — morfoloogilise ja süntaktilise analüüsi põhjal, C — morfoloogilise, süntaktilise ja semantilise analüüsi kaudu. Kõige sügavam on tasand D, kus sisendlause tähendus esitatakse niisuguses universaalses semantilises koodis, mille puhul keeltevahelised erinevused on täielikult kadunud. Sel juhul ei ole vaja mingit erilist üleminekut ühelt keelelt teisele: sellestsamast struktuurist, mille saime lähtelause analüüsi tulemusena, võime alustada väljundkeele vastava lause või lausete sünteesimist.

Viimati nimetatu on tõlkimise ideaal, kuivõrd selle tegevuse tuum on ühe keele teksti mõtte edasiandmine teises keeles või teistes keeltes. Kuid ehkki tõlkeprotsessi niisuguse mudeli idee esitati juba 1960-ndate aastate alguses (esmakordselt ilmselt 1961. aastal Moskvas toimunud masintõlke ja infotöötluse konverentsil), ei ole ette näha selle taseme saavutamist lähemas tulevikus.

Ühelt keelelt teisele ülemineku tasandid on tegelikult terved kihid. Näiteks on süsteeme, mis peamiselt morfoloogilisele analüüsile toetudes kasutavad erineval määral süntaktilise analüüsi elemente. Viimasel ajal loodud ning loomisel olevad süsteemid püüavad võimalikult täieliku süntaktilise analüüsi juures erineval määral kasutada semantilise analüüsi andmeid.

Masintõlkesüsteemi tööst kõneldes tuleb lahus hoida ühelt poolt otseseid tekstianalüüsi või -sünteesi protseduure ja teiselt poolt keelematerjali — sõnade, sõnaühendite ja lausete — struktuuri formaalse esitamise süsteemi, mida kasutatakse nende protseduuride



Joon. 5

käigus. Näiteks morfoloogilise analüüsi tulemuseks on vastava lause morfoloogiline esitus, milles on näidatud lauses sisalduvate sõnavormide liigendumist morfoloogilisteks elementideks — tüvedeks, sufiksiteks, tunnusteks, muutelõppudeks. Selle esituse kirjapanemiseks on vaja formaalsete vahendite süsteemi, mis on suhteliselt sõltumatu analüüsi enese käigust. Niisamuti räägitakse lause süntaktilisest esitusest ja semantilisest esitusest. Esimene neist on süntaktilise analüüsi, teine semantilise analüüsi tulemus. Kõnealune eristus ning keelematerjali esituse ja analüüsi-sünteesialgoritmide sõltumatus on olulised selleks, et üht ja sedasama esitusviisi saaks kasutada nii analüüsis kui ka sünteesis. Kui näiteks on välja töötatud formaalne esitusviis eesti keele lausete süntaktilise struktuuri esitamiseks, siis saame seda kasutada niihästi tõlkides eesti keelest mõnda teise keelde (seega analüüsi käigus) kui ka tõlkides mõnest teisest keelest eesti keelde (seega

eesti keele lausete sünteesi käigus). Pealegi on ainult tõlkimisprotsessi niisuguse korraldusega võimalik saavutada, et masintõlkesüsteem ei oleks orienteeritud ainult kindla keeltepaari vahel tõlkimiseks ning et antud keelest mingisse kolmandasse keelde tõlkimiseks ei oleks vaja välja töötada hoopis uusi esitusviise ja analüüsialgoritme.

Vastavalt kõnealusele eristusele võib rääkida ühelt poolt masintõlkesüsteemi lingvistilisest varustusest ehk keelevarast, mis haarab formaalsete esitusvahendite süsteemi lausete struktuuri kirjeldamiseks kõikidel tasanditel, ja teiselt poolt tema algoritmvarustusest ehk algoritmvarast, millesse kuuluvad otsesed analüüsi, sünteesi ja ühelt keelelt teisele ülemineku algoritmid ning protseduurid.

Käesolevas raamatus on kõne all eelkõige masintõlkesüsteemide keelevara. Analüüsi- ja sünteesiprotseduuridest vaatleme ainult nende ehituspõhimõtteid.

Käesolevas peatükis kirjeldame lähemalt morfoloogilist ja süntaktilist analüüsi-sünteesi ja nendel etappidel tarvisminevat keelevara, teiste sõnadega masintõlkesüsteemi morfoloogilist ja süntaktilist komponenti. Et lausete tähendust esiletoovat semantilist analüüsi kasutatakse masintõlkesüsteemides piiratud ulatuses, jätame nende probleemide käsitluse järgmisse peatükki — seda enam, et keelt mõistvates süsteemides on just semantikaprobleemid keskse kohal.

Morfoloogilise analüüsi ja sünteesi puhul tuleb kõigepealt arvestada, et just morfoloogilise ehituse — sõnade morfoloogilise vormistuse — osas võivad keeled kõige rohkem üksteisest erineda. Meile lähemalt tuntud keeled, eelkõige Euroopas kõneldavad keeled erinevad üksteisest õnneks suhteliselt vähe, kuid siingi on mõnedes aspektides erinevused üsna märgatavad.

Näiteks eesti keeles on nimisõnade ja omadussõnade puhul üheks olulisemaks morfoloogiliseks kategooriaks kääne. Eelkõige sõna käände põhjal saame aru, missuguse teise sõnaga on antud nimi- või omadussõna lauses seotud, ja ka seda, missuguse sisulise seose kaudu. Nimisõna ja temaga seotud omadussõna on üldreeglina samas käändes: *uues majas, uuel majal, uuest majast;* et fraasid *rääkis tüdrukule* ja *rääkis tüdrukust* kirjeldavad hoopis erinevaid situatsioone, seda mõistame samuti nimisõna *tüdruk* erinevate käänete põhjal.

Ka vene või saksa keeles muutuvad nimi- ja omadussõnad käänetes, kuid käändeid on märksa vähem kui eesti keeles, samal ajal on märksa rohkem eessõnu ehk prepositsioone, mis täidavad põhimõtteliselt samu funktsioone kui käänded. Inglise ja prantsuse keeles käänded praktiliselt puuduvad. Sõnadevahelised seosed, mida meil väljendatakse käänetega, antakse nendes keeltes edasi pre- ja postpositsioonidega, aga ka sõnade järjekorraga lauses. Seetõttu ongi üleminek keelelt keelele morfoloogilisel tasandil niivõrd keeruline. Morfoloogiakomponent on tänapäeva masintõlkesüsteemides va-

hendav lüli teksti graafilise esituse ning tema süntaktilise esituse kui põhilise, aga osalt ka semantilise esituse vahel. Morfoloogiline analüüs peab seetõttu iga sõna puhul esile tooma need andmed, mis on vajalikud süntaktiliseks analüüsiks või kannavad sisulist informatsiooni, mida väljundkeeles on vaja vastavate vahenditega edasi anda. Nagu juba osutasime, täidavad käänded eelkõige süntaktilist funktsiooni. Teiselt poolt tegusõnade aeg (olevik, minevik jt.) ja kõneviis (kindel, tingiv jt.) ei täida lauses mitte sõnu siduvat funktsiooni, vaid omavad iseseisvat semantilist sisu.

Morfoloogiakomponendi sisendiks analüüsietapil on niisiis lause kui sõnavormide järjend, väljundiks samade sõnade loend, kus iga sõna juures on vastavate kategooriate termineis esile toodud tema morfoloogilised tunnused. Näiteks eesti keele lause *Poisid lidusid hirmunult esikusse* peaks morfoloogilise analüüsi järel välja nägema järgmiselt:

poiss (nimisõna; mitmus; nimetav kääne)
liduma (tegusõna; kindel kv.; lihtminevik; mitmus; 6. pööre)
hirmunult (määrsõna)
esik (nimisõna; ainsus; sisseütlev kääne)

Morfoloogilise analüüsi peamised vahendid on: 1) sõnatüvede sõnastik; 2) nende morfoloogiliste vormielementide (tunnuste lõppude, eesliidete jne.) loendid, mis kindlates kombinatsioonides sõnatüvedega võivad vastavas keeles moodustada konkreetseid sõnavorme; 3) andmed selle kohta, missuguste sisuliste morfoloogiliste kategooriate abil tõlgendada üht või teist sõnatüve ja morfoloogiliste vormielementide ühendit.

Sõnatüvede sõnastikus on iga sõnatüve juures sõnaliik ning mitmesugused andmed, mis on vajalikud antud tüvevormi äratundmiseks (analüüsis) või sellest tüvevormist lähtudes nõutava tervikliku morfoloogilise vormi sünteesimiseks.

Kui sõnatüvi võib esineda mitmel eri kujul, siis tuleb need kõik sõnastikus üles lugeda. Eesti keeles on mitme erineva tüvekuju esinemine teatavasti sagedane nähtus (meenutame meie astmevaheldust: *hakkama — hakata, viga — vea, vend — venna*). Tuleb kohe öelda, et see, mida niisuguses automaatanalüüsi ja -sünteesi jaoks koostatud sõnastikus mõistetakse sõnatüve all, ei pruugi ühtida sõnatüve mõistega ranges keeleteaduslikus käsitluses. Meie juhtumil on aluseks eelkõige praktiline kriteerium: sõnastikus tuleb esitada kõik need vormid (sõnaosad), mis ei ole teistest vormidest küllalt üldiste ja lihtsate, seejuures täiesti formaalsete reeglite abil tuletatavad.

Püüame masintõlkesüsteemi morfoloogiakomponendi ehitust ja tööd illustreerida eesti keele näidete najal. Eesti keele käändsõnade (nimi-, omadus-, arv- ja asesõnad) kõigi käändevormide kirjeldamiseks on vaja sõnastikus esitada seitse või rohkemgi erinevat vormi, s. t. eri tüve: ainsuse nimetava, omastava, osastava ja lühikese sisseütleva

vormid, mitmuse omastava ja osastava vormid (viimaseid võib olla mitu) ning *i*-mitmuse tüvi. Need ongi vormid, mida ei saa üksteisest küllalt üldiste reeglite abil tuletada või ära tunda. Erand on mitmuse omastav, mida saab võrdlemisi regulaarselt moodustada ainsuse osastavast *(pesa—pesade, puu—puude, katust—katuste)*, kuid kui arvestada, et selle vormi abil moodustatakse kõik mitmuse käänded alates sisseütlevast, siis on ilmselt otstarbekam ta sõnastikus esitada, selle asemel, et teda alati uuesti moodustama hakata. Loetletud vormid tuleb kindlal viisil tähistada, näiteks nummerdada: TÜVI 1, TÜVI 2 jne., osutades iga vormi puhul ühtlasi, missuguse sõna vormiga on tegemist. Näiteks: *poisi:* N TÜVI 2 *poiss* tähendab, et vorm *poisi* on sõna *poiss* tüvevariant number 2 ning tema sõnaliik on nimisõna (N). Toome mõned näited.

raamat:	N	TÜVI 1	*raamat*
raamatu:	N	TÜVI 2	*raamat*
raamatut:	N	TÜVI 3	*raamat*
raamatute:	N	TÜVI 5	*raamat*
raamatuid:	N	TÜVI 6	*raamat*
raamatut:	N	TÜVI 7	*raamat*
pesa:	N	TÜVI 1	*pesa*
		TÜVI 2	*pesa*
		TÜVI 3	*pesa*
pessa:	N	TÜVI 4	*pesa*
pesade:	N	TÜVI 5	*pesa*
pesi:	N	TÜVI 6	*pesa*
		TÜVI 7	*pesa*
pesasid:	N	TÜVI 7	*pesa*
pikk:	OM	TÜVI 1	*pikk*
pika:	OM	TÜVI 2	*pikk*
pikka:	OM	TÜVI 3	*pikk*
		TÜVI 4	*pikk*
pikkade:	OM	TÜVI 5	*pikk*
pikki:	OM	TÜVI 6	*pikk*
piki:	OM	TÜVI 7	*pikk*

Neis loendites tähistab N nimisõnade, T tegusõnade, OM omadussõnade ja M määrsõnade klassi. Nagu näeme, võivad mõnel sõnal mitu tüvekuju vormiliselt ühtida, pealegi ei pruugi igal sõnal olla kõiki võimalikke tüvevorme. Näiteks puudub sõnal *raamat* TÜVI 4 (lühike sisseütlev).

Lisaks tüvede sõnastikule on morfoloogiakomponendis morfoloogiliste vormielementide loendid. Meie näite puhul on selleks käändelõppude loend: *d, sse, s, st, le* jne. Konkreetsete vormide kirjeldamiseks on reeglid, mis ütlevad, missuguse morfoloogilise vormiga on tegemist, kui on antud mingi tüve kombinatsioon mingi käändelõpuga (võib olla ka tüvevorm üksi). Näiteks:

1) TÜVI 1 → ainsus; nimetav kääne
2) TÜVI 2 → ainsus; omastav kääne
3) TÜVI 3 → ainsus; osastav kääne
4) TÜVI 4 → ainsus; sisseütlev kääne
5) TÜVI 2+*sse* → ainsus; sisseütlev kääne
6) TÜVI 2+*d* → mitmus; nimetav kääne
7) TÜVI 5 → mitmus; omastav kääne
8) TÜVI 6 → mitmus; osastav kääne

9) TÜVI 5+*lt* → mitmus; alaltütlev kääne
10) TÜVI 7+*lt* → mitmus; alaltütlev kääne

jne.

Teine peamine muutuvate sõnade klass eesti keeles on pöördsõnad, millel on muidugi hoopis teistsugune vormistik ehk vormide paradigma kui käändsõnadel, kuid selle kirjelduspõhimõtted on samasugused. On tarvis välja selgitada tüvevormide hulk, millest vastavaid tunnuseid ja lõppe lisades on võimalik moodustada kõik tegusõnavormid.

Eesti keeles ei ole tegusõnade eri tüvevariantide arv suurem kui nimisõnadel, pigem väiksem. Üldreeglina on vaja teada vaid *ma*-infinitiivi tüve (TÜVI 1), *da*-infinitiivi tüve (TÜVI 2), kindla kõneviisi oleviku tüve (TÜVI 3) ning umbisikulise tegumoe oleviku *(võetakse, saetakse)* tüve (TÜVI 4). Kuid morfoloogiline vormistik tervikuna on tegusõnadel märksa keerulisem kui käändsõnadel. Kui käändsõnade ainsad äratundmist vajavad kategooriad on kääne, arv (ainsus ja mitmus) ja omadussõnadel veel võrdlusaste (keskvõrre, ülivõrre), siis tegusõnadel on kategooriaid terve hulk. Kõigepealt on pöördeliste vormide puhul vaja ära tunda tegumood (isikuline, umbisikuline), kõneviis (kindel, tingiv, käskiv, kaudne), aeg (olevik, liht-, täis-, enneminevik), arv (ainsus, mitmus) ning lõpuks pööre ise. Seejuures sisaldab iga pöördeline vorm sisuliselt kõikide loetletud kategooriate mingit esindajat — ta on mingis tegumoes, kõneviisis, ajas, arvus, pöördes. Lisaks pöördelistele vormidele on tegusõnadel veel rida käändelisi vorme: infinitiivid (*ma*-, *da*-infinitiiv) ja nende mitmesugused vormid (*mas*-, *mast*-, *maks*-, *mata*-, *des*-vorm) ning kesksõnad (*nud*-, *tud*-, *v*-, *tav*-kesksõna).

Niisiis on tegusõnadel morfoloogiliste vormielementide (tunnuste ja lõppude) hulk, mis vastavas sõnastikus tuleb loetleda, märksa ulatuslikum kui käändsõnadel; ka reeglid, mille põhjal analüüsitavas tekstis erinevad sõnavormid ära tuntakse või sünteesitava teksti sõnavormid moodustatakse, on siin tunduvalt keerulisemad. Detailidesse laskumata toome siingi mõned näited.

hüppa	T	TÜVI 1	*hüppama*
		TÜVI 3	*hüppama*
hüpa	T	TÜVI 2	*hüppama*
		TÜVI 4	*hüppama*
kukku	T	TÜVI 1	*kukkuma*
		TÜVI 2	*kukkuma*
kuku	T	TÜVI 3	*kukkuma*
		TÜVI 4	*kukkuma*
võt	T	TÜVI 1	*võtma*
		TÜVI 2	*võtma*
võta	T	TÜVI 3	*võtma*
võe	T	TÜVI 4	*võtma*
lidu	T	TÜVI 1	*liduma*
		TÜVI 2	*liduma*
		TÜVI 3	*liduma*
		TÜVI 4	*liduma*

1) TÜVI 3+*n*
→ isikuline tegum.; kindel kv.; olevik; ainsus; 1. pööre

2) TÜVI 3+*ksi*+*me*
→ isikuline tegum.; tingiv kv.; olevik; mitmus; 4. pööre
3) TÜVI 1+*si*+*d*
→ isikuline tegum.; kindel kv.; lihtminevik; ainsus; 2. pööre
→ isikuline tegum.; kindel kv.; lihtminevik; mitmus; 6. pööre

4) TÜVI 1+*ma* → MA-INF
5) TÜVI 2+*des* → DES
6) TÜVI 2+*tes* → DES
7) TÜVI 2+*es* → DES

Viimases kolmes reeglis on DES *des*-vormi üldistatud tähistuseks.

Tegusõnade puhul on eesti keeles, nii nagu paljudes teisteski keeltes, omaette probleemiks liitaegade vormid, mis moodustuvad kahest sõnavormist — sõna *olema* vastavast vormist ja *nud*- või *tud*-kesksõnast *(olen hüpanud, olid hüpanud, oldi hüpatud)*. Morfoloogilise analüüsi esimesel etapil analüüsitakse vastavaid sõnavorme üldiste reeglite järgi. Alles seejärel, kui kõik üksikud sõnavormid on identifitseeritud, s. o. nende ehitus morfoloogiliste kategooriate termineis kindlaks tehtud, hakatakse eriprotseduuride abil kontrollima mitmesõnaliste vormide olemasolu lauses. Liitaegade kõrval on teiseks regulaarseks mitmesõnaliste vormide tüübiks tegusõnade eitavad vormid *(ei hüppa, ei hüpanud, ei olevat hüpanud)*.

Morfoloogilise analüüsi protseduur ise on suhteliselt lihtne. Konkreetseid algoritme on eri keelte jaoks välja töötatud palju, kuid nad on üksteisega üldjoontes sarnased. Kõigepealt kontrollitakse, kas antud vorm tervikuna sisaldub tüvede sõnastikus. Kui ta ei sisaldu, siis hakatakse sõna lõpust ühekaupa tähti eemaldama ja kontrollitakse iga kord, kas sõnaosa, mis pärast vastava tähe eemaldamist järele jäi, leidub tüvede sõnastikus. Kui ei leidu, eemaldatakse järgmine täht. Kui leidub, kontrollitakse, kas sõna lõpust äravõetud sõnaosa vastab mingile üksusele morfoloogiliste elementide loendis või nende mingile kombinatsioonile. Kui niisugune element või nende kombinatsioon leitakse, otsitakse reegel või reeglid, mis ütlevad, missuguse morfoloogilise vormiga on tegemist. Reeglite abil konstrueeritud morfoloogiliste kategooriate järjend ongi sõna esitus ehk kirje morfoloogilisel tasandil. Lause morfoloogiliseks esituseks on aga selles sisaldunud sõnavormide morfoloogiliste kirjete loend.

Eespool juba tõime näite lause morfoloogilisest esitusest. Täielikkuse huvides lisame sellele teise, pisut keerulisema lause *Poiss oleks peaaegu katuselt tänavale kukkunud:*

poiss	(N; ainsus; nimetav kääne)
kukkuma	(T; isikuline tegum.; tingiv kv.; täisminevik; ainsus; 3. pööre)
peaaegu	(M)
katus	(N; ainsus; alaltütlev kääne)
tänav	(N; ainsus; alaleütlev kääne)

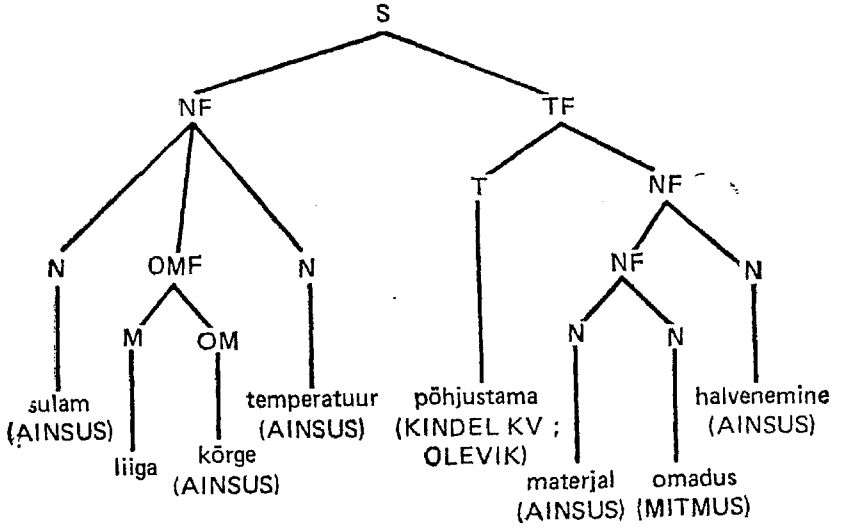
Olgu lisatud, et kõik morfoloogilised kategooriad, mida me selguse

huvides sõnadega nimetasime, on arvutis mäluruumi säästmiseks esitatud enamasti numbriliste kooditähistustega. Ka sõnatüved võib esitada kodeeritult, näiteks sobib sõnatüve tähistajaks hästi tema järjenumber tüvede sõnastikus.

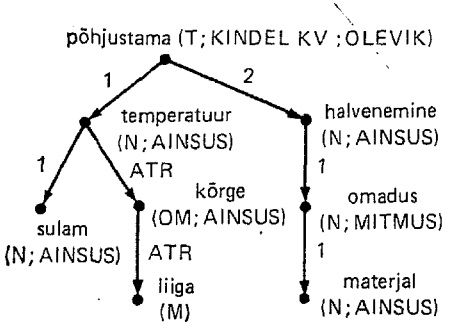
Süntaktiline analüüs on juba märksa keerulisem lausete automaatanalüüsi protseduur. Märksa keerulisem on ka vastav keelevara — lause süntaktilise struktuuri esitamiseks kasutatav formaalsete vahendite süsteem. Samuti valitseb süntaktilise analüüsi meetodites ja esitusviisides hoopis suurem mitmekesisus.

Lause süntaktilise struktuuri kirjelduse ehk süntaktilise esituse all mõistetakse lause niisugust kujutust, kus on formaalselt näidatud seosed lauses sisalduvate sõnavormide vahel. Seosed määratakse morfoloogilise analüüsi tulemusel leitud üksuste vahel informatsiooni põhjal, millega nad varustati analüüsi käigus. Seega võime öelda, et masintõlkesüsteemi süntaksi komponendi sisendiks analüüsi puhul on lause morfoloogiline esitus ja väljundiks süntaktiline esitus. Sünteesi puhul on vastupidi: süntaksikomponendi sisendiks on lause süntaktiline esitus, mis on saadud tõlgitava keele lause süntaktilisest esitusest, väljundiks aga morfoloogiline esitus.

Lause süntaktilise struktuuri esitusviisid võib kogu nende mitmekesisuse juures taandada kahele



Joon. 6



Joon. 7

põhitüübile — seda mitte ainult masintõlkesüsteemides, vaid ka teistes keeletöötlussüsteemides, aga ka keele mitterakenduslikes teoreetilistes käsitlustes: need on fraasistruktuuriesitus ja sõltuvusesitus. Mida nad endast kujutavad, seda on kõige parem selgitada konkreetse näite varal. Vaatame, kuidas näeb välja lause *Sulami liiga kõrge temperatuur põhjustab materjali omaduste halvenemise* süntaktilise struktuuri kirjeldus fraasistruktuuriesituses (joon. 6) ja sõltuvusesituses (joon. 7).

Nagu näeme, on esitusviisid väliselt võrdlemisi erinevad, ehkki aluseks olev põhiidee on üks ja sama: otsustada, missugused sõnad lauses kuuluvad kokku. Fraasistruktuuriesitus teeb seda, tuues esile lause hierarhilise jagunemise fraasideks, kus fraasi moodustab teatud sõna koos oma laienditega. Fraasid liigitatakse vastavalt põhisõna liigile: nimisõnafraas NF, tegusõnafraas TF, omadussõnafraas OMF, määrsõnafraas MSF jne. Altpoolt ülespoole liikuma hakates võime niisuguse struktuuri põhjal kindlaks teha, missugused sõnad kuuluvad kokku ja missugusel tasandil. Näiteks võime öelda, et sõnad *liiga* ja *kõrge* kuuluvad kokku, moodustades omadussõnafraasi; et *sulam* kuulub ühte sõnaga *temperatuur*, moodustades fraasi alles koos temaga, mitte aga vahetult järgnevate sõnadega *liiga kõrge;* jne. Lause tervikuna jaguneb kõigepealt nimisõnafraasiks ja tegusõnafraasiks, mis enam-vähem vastavad traditsioonilise grammatika alus- ja öeldisrühmale.

Erinevalt fraasistruktuuriesitusest ei ole sõltuvusesituse eesmärk kujutada lause järkjärgulist jagunemist väiksemateks osadeks, vaid näidata vahetult sõnade vahel kehtivaid sõltuvusi. Seepärast pole sõltuvusesituses omaette üksustena niisuguseid kategooriaid nagu fraasistruktuuriesituses. Seosed viivad sõna juurest sõna juurde, seejuures on nad alati kindla suunaga. Skeemil suundub nool sõltuvuse pealiikme juurest sõltuva liikme juurde. Traditsiooniline grammatika tunneb teatavasti kaht lause pealiiget — alust ja öeldist. Sõltuvusesituses opereeritakse ainult ühe pealiikmega, milleks on üldreeglina lause öeldiseks olev tegusõna; aluseks olev nimisõna on vaid üks selle laiendeid. Sellest, missugune tegusõna on öeldiseks, sõltub ju vormiliselt ka see, kas lauses üldse saab esineda alus, ja kui saab, siis ka see, missugused sõnad selles positsioonis võivad esineda.

Sõnadevahelised seosed on alati kuidagi tähistatud või nimetatud. Tähistused iseloomustavad sõltuvuse sisulist tüüpi. Näiteks on eeltoodud sõltuvusskeemis sõna *põhjustama* seosed sõnadega *temperatuur* ja *halvenemine* erinevalt tähistatud, seega on need laiendid oma põhisõnaga erinevates seostes.

Fraasistruktuuriesitusel ja sõltuvusesitusel on kummalgi omad eelised ja puudused. Neisse lähemalt süvenemata osutame, et viimasel ajal on valitsevaks muutunud sõltuvusesitus. Esiteks on ta märksa kompaktsem, milles võib veenduda mõlema skeemi visuaalsel võrdlemisel. Teiseks on siin otseselt ja selgelt näidatud sisulised sõltuvused sõnade vahel, need sõltuvused on aga osutunud peamiseks lause süntaktilist struktuuri iseloomustavaks informatsiooniliigiks. Fraasistruktuuriesituse puhul selguvad sõnadevahelised seosed vaid kaudselt, pealegi ei ole alati üheselt selge sõltuvuse suund, kuna sõltuvuste sisu pole üldse eristatud. Lõpuks osutub sõltuvusesitus märksa mugavamaks lause süntaktilise ja semantilise struktuuri seostamisel, mille juurde jõuame järgmises peatükis.

Järgnevalt masintõlkesüsteemi süntaksikomponendist rääkides vaatleme ainult sõltuvusesituse kasutamisega seotud probleeme.

Nagu öeldud, on süntaksikomponendi lähtematerjaliks lause morfoloogiline esitus — sõnade rida, kus iga sõna juures on esitatud teda iseloomustav morfoloogiline informatsioon. Et sellest materjalist lähtudes konstrueerida lause süntaktiline esitus, kasutatakse süntaksikomponendis esiteks sõnastikku ja teiseks suurt hulka mitmesuguse funktsiooniga reegleid.

Eri süsteemides on nii sõnastikus esitatav materjal kui ka reeglite süsteem üpris erinevad. Püüame kirjeldada mingit üldistatud, «keskmist» süntaksikomponenti. Allpool kirjeldatava kõige lähemaks eeskujuks on prantsuse-vene masintõlkesüsteem, mida töötatakse välja Moskvas instituudi «Informelektro» teoreetilise ja eksperimentaallingvistika sektoris.[32]

Sõnastikus tuleb iga sõna juures esile tuua tema püsivad süntaktilised ja semantilised omadused — need, mis ei *muutu* lausest lausesse ja kasutusest kasutusse. Peale üksiksõnade on sõnastikus omaette üksustena esitatud ka fraseoloogilised ühendid, mis lause süntaktilises struktuuris funktsioneerivad kui tähenduslikud tervikud (*jalga laskma, varvast viskama* jt.). Iga sõna või fraseoloogilise ühendi juures on vaja esitada vähemalt järgmised andmed (ehk, nagu öeldakse: sõnastikuartikkel jaguneb järgmisteks tsoonideks): 1) sõna tõlge teise keelde, 2) sõna semantilised tunnused, 3) sõna süntaktilised tunnused, 4) sõna sõltuvusmall.

Sõna semantilised tunnused on üldistatud kategooriad, mille abil iseloomustatakse sõna tähendust, nagu TEGEVUS, SEISUND,

[33] Апресян Ю. Д. Лингвистическое обеспечение в системе автоматического перевода третьего поколения. М., АН СССР. Научный совет по комплексной проблеме «Кибернетика». Препринт. 1978.

PROTSESS, ESE, ELUSOLEND, INIMENE, INIMRÜHM jne. Näiteks sõnal *jooksma* on üks semantilisi tunnuseid TEGEVUS, sõnadel *poiss, mees* ja *laps* — INIMENE, sõnadel *raamat* või *maja* — ESE jne. Semantilisi tunnuseid läheb vaja selleks, et analüüsi käigus välistada grammatiliselt võimalikke, kuid sisuliselt võimatuid sõnaühendeid (lähemalt räägime sellest allpool).

Süntaktilised tunnused sarnanevad väliselt semantilistega, kuid nad ei iseloomusta sõnade tähendust, vaid nende võimalikke süntaktilisi rolle lauses. Süntaktilised tunnused omistatakse sõnale, lähtudes tema võimest moodustada sõnaühendeid teiste sõnatüüpidega. Ka sõnaliigid on selles mõttes süntaktilised tunnused. Kuid keele formaalses kirjelduses on tarvis märksa peenemat sõnade liigitust, kui seda on traditsiooniline jaotus sõnaliikideks. On võrdlemisi tavaline, et sõnaliikide asemel kasutatakse 30—40 süntaktilist sõnaklassi. Eriti komplitseerituks osutuvad «väikeste» sõnaliikide, nagu asesõnade, sidesõnade jt., jaotused. Omaette klassidena tuleb vaadelda ka kirjavahemärke (punkt, koma, koolon jt.), matemaatilisi avaldisi, kui need on lükitud otse lausesse, jne.

Sõltuvusmall ei pruugi sisalduda kõigi sõnade kirjeldustes, vaid antakse üksnes sõnadel, mille juurde võivad kuuluda kindlat tüüpi sõltuvad üksused, laiendid. Niisuguse sõna esinemine lauses «ennustab» ka tema laienditeks olevate teiste sõnade (vormide) võimalikku või isegi kohustuslikku olemasolu samas lauses. Näiteks sõna *ostma* põhjal võime ennustada, et lauses on olemas ka sõnad, mis väljendavad seda, kes ostab, mida või mille ostab, kellelt ostab, missuguse hinna eest ostab (kui mitte kõik korraga, siis mõned neist kindlasti). Seejuures määrab sõna *ostma* nii laiendite olemasolu kui ka nende võimaliku morfoloogilise vormi, antud juhul vastavate nimisõnade käände. Sõltuvusmall esitabki informatsiooni sõnade võimalike või kohustuslike laiendite kohta lauses. Sõna võimet nõuda teatud kindlaid laiendeid kirjeldab tema valents (termin on moodustatud ilmse analoogia põhjal keemiaga). Sõltuvusmallid on olemas peaaegu kõikidel tegusõnadel, paljudel omadus- ja määrsõnadel, aga ka nimisõnadel. Näiteks nimisõna *temperatuur* puhul esineb harilikult laiend — nimisõna omastavas käändes —, mis osutab, kelle või mille temperatuurist on jutt *(haige temperatuur, sulami temperatuur)*. Sõltuvusmallis ei loetleta laiendeid, mis ei sisult ega vormilt pole spetsiifilised antud sõnale, vaid võivad esineda koos paljude sõnadega; tegusõnadel on niisugusteks näiteks tegevuskohta, aega, viisi jt. aspekte iseloomustavad laiendid.

Sõna sõltuvusmallis on sõna iga võimaliku laiendi kohta esitatud järgmine informatsioon: 1) antud laiendiks sobiva sõna võimalikud semantilised tunnused; 2) laiendi süntaktilised tunnused ja võimalikud morfoloogilised avaldumisvormid (nimisõna kindlas käändes,

infinitiiv, kesksõna, kõrvallause); 3) sõltuvusseose sisuline tüüp, mis seob põhisõna antud laiendiga.

Mis puutub sõltuvusseose sisulistesse tüüpidesse, siis eri süsteemides kasutatakse vägagi erinevaid seoste hulki sõltuvalt sellest, kui peenelt peetakse vajalikuks sõnadevahelisi seoseid sisuliselt eristada. Eespool viidatud prantsuse-vene masintõlkesüsteemis kasutatakse süntaktilise struktuuri lõppesituses üheksat seost, ehkki vahepealsel etapil, lause analüüsi käigus, esineb kuni nelikümmend süntaktilise seose liiki. Mainitud üheksast seosest on esimesed kuus aktandiseosed, mis seovad tegusõnu, aga ka teisi sõnaliike nende rektsiooniliste laienditega. Näiteks seob seos nr. 1 tegevust või protsessi selle subjekti või kandjaga, seos nr. 2 selle objektiga, seos nr. 3 kaudobjektiga, seos nr. 4 selle vahendiga, seos nr. 5 tegevuse eesmärgi või otstarbega, seos nr. 6 tegevusajaga jne. Toome illustratsiooniks mõned näited: *poiss* $\overset{1}{\rightarrow}$ *jooksis, poisi* $\overset{1}{\rightarrow}$ *jooksmine, puude* $\overset{2}{\rightarrow}$ *lõhkumine, kinkima* $\overset{2}{\rightarrow}$ *raamatut, kinkima* $\overset{3}{\rightarrow}$ *tüdrukule, lõhkuma* $\overset{4}{\rightarrow}$ *kirvega, kasvatama* $\overset{5}{\rightarrow}$ *müügiks, komandeerima* $\overset{6}{\rightarrow}$ *(kaheks) nädalaks.*

Eespool kirjeldatu haarab tegelikult vaid sõna sõltuvusmalli kondikava. Loetletud andmetele lisanduvad andmed üksikute laiendite kohustuslikkuse või mittekohustuslikkuse kohta antud sõnal, samuti võivad kehtida vägagi keerulised kitsendused erinevate laiendite, eriti nende kindlate vormivariantide samaaegse esinemise suhtes. Näiteks sõnast *käskima* võib sõltuda nimisõna alalütlevas käändes: *käskis poisil (teha midagi)*, sedasama valentsi võib väljendada ka nimisõna osastavas käändes: *käskis poissi.* Samuti võib sõnal *käskima* olla laiend, mis esitab käsu sisu ja mis võib esineda *da*-tegevusnimelise sõnaühendina *(käskis teha midagi)* või *et*-kõrvallausena *(käskis, et tehtaks midagi).* Nende kahe laienditüübi samaaegsel kasutamisel kehtib aga kitsendus, et kui esimene valents on täidetud nimisõnaga alalütlevas käändes, ei saa teise valentsi täitmiseks kasutada kõrvallauset: ei saa öelda ... *käskis poisil, et see teeks midagi.* Saab öelda: *käskis poisil teha midagi* või *käskis poissi, et see teeks midagi.*

Toome mõned näited sõnade esitusest süntaksikomponendi sõnastikus (sõnade tõlkevasted jätame esitamata).

temperatuur
1. Semantilised tunnused: MÕÕDE; MÕÕTE VÄÄRTUS.
2. Süntaktiline tunnus: N.
3. Sõltuvusmall:
   1. aktant:
   1) semantilised tunnused: ESE; ELUSOLEND; AINE; AJAVAHEMIK,
   2) vorm: N + omastav kääne.

Selgituseks. Esiteks, kaks erinevat semantilist tunnust osutavad, et sõna *temperatuur* kasutatakse kord vastava mõõtme kui niisuguse nimetusena *(keha temperatuur*

oli 37 kraadi, s. o. mõõtme 'temperatuur' väärtus oli 37 kraadi), kord aga nimetamaks seda mõõdet koos konkreetse väärtusega: *temperatuur langeb* (=mõõtme 'temperatuur' väärtus väheneb).

Teiseks, sõltuvusmallis on andmed vaid ühe tüüpilise laiendi kohta, mis on seotud sõnaga *temperatuur* 1. aktandiseose kaudu. Seejuures võivad selle laiendina esineda sõnad väljendada esemeid: *kivi temperatuur;* elusolendeid: *lapse temperatuur;* ainet: *vee temperatuur;* ajavahemikku: *jaanuarikuu (keskmine) temperatuur.* Vormiliselt esineb see laiend nimisõnana omastavas käändes.

käskima
1. Semantiline tunnus: TEGEVUS.
2. Süntaktiline tunnus: T.
3. Sõltuvusmall:
   1. aktant:
   (kes?)
   1) semantiline tunnus: INIMENE,
   2) vorm: N + nimetav kääne;
   2. aktant:
   (keda?)
   1) semantilised tunnused: INIMENE; LOOM,
   2) vorm:
      a) N + osastav kääne
      b) N + alalütlev kääne;
   3. aktant:
   (mida?)
   1) semantiline tunnus: TEGEVUS,
   2) vorm:
      a) *da*-infinitiiv
      b) *et* + kõrvallause.

Kuivõrd sõnast *käskima* oli äsja juttu, siis täiendavad seletused pole vajalikud. Niisamuti peaksid eelnenud seletuste põhjal olema arusaadavad järgmised näited.

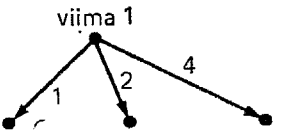
viima 1
1. Semantilised tunnused: TEGEVUS; LIIKUMINE.
2. Süntaktiline tunnus: T.
3. Sõltuvusmall:
   1. aktant:
   (kes?)
   1) semantilised tunnused: INIMENE; LOOM,
   2) vorm: N + nimetav kääne;
   2. aktant:
   (mida? mille?)
   1) semantilised tunnused: ESE; ELUSOLEND; AINE,
   2) vorm:
      a) N + ainsus + + omastav k.
      b) N + mitmus + + nimetav k.
      c) N + osastav k.;
   3. aktant:
   (millega?)
   1) semantilised tunnused: VAHEND; ESE,
   2) vorm:
      N + kaasaütlev k.
      N + omastav k. + *abil.*

Vorm *viima 1* osutab, et toodud sõnastikuartikkel kirjeldab sõna *viima* selle ühes kindlas tähenduses, ruumilise liikumise tähenduses.

õu
1. Semantiline tunnus: KOHT.
2. Süntaktiline tunnus: N.
3. Sõltuvusmall: puudub.

isa

1. Semantiline tunnus: INIMENE.
2. Süntaktiline tunnus: N.
3. Sõltuvusmall:
   1. aktant:
      1) semantiline tunnus: INIMENE,
      2) vorm: N + omastav k.



Joon. 8

Rõhutame, et need on üksnes illustratiivsed näited, vormistatud ülevaatlikkuse huvides võimalikult lihtsatena. Nagu märgitakse J. Apresjani jt. eespool viidatud töös, võtab näiteks sõna *temperatuur* kõikvõimalikke kasutusi ja muid aspekte haarav kirjeldus (vene keeles) enda alla umbes 10 lehekülge masinkirjateksti.[34]

Sõnastikuartikli niisugusel ülesehitusel on selge eesmärk. Sõna semantilised ja süntaktilised tunnused on vajalikud esmajoones sel puhul, kui antud sõna täidab mõne teise sõna laiendi rolli. Sõltuvusmalli põhjal võib aga lauses leida sõnad, mis sobivad antud sõna laienditeks. Näiteks sõna *viima 1* sõltuvusmall määrab järgmise struktuurifragmendi (joon. 8).

Sõnastikuartikli aktantide juures esitatud semantilised ja vormilised tunnused osutavad, missuguseid tingimusi peavad rahuldama sõnad, mis võivad täita vastavate aktantide rolli. See tähendab, nad osutavad, missuguseid teisi sõnavorme tasub otsida lauses, kus esineb antud sõna (näiteks *viima 1*).

Pöördume nüüd sõnastiku juurest süntaksikomponendi reeglite juurde. Reeglid võib nende funktsiooni järgi jagada kahte liiki: süntaktilisi seoseid kehtestavad reeglid ja filtrid.

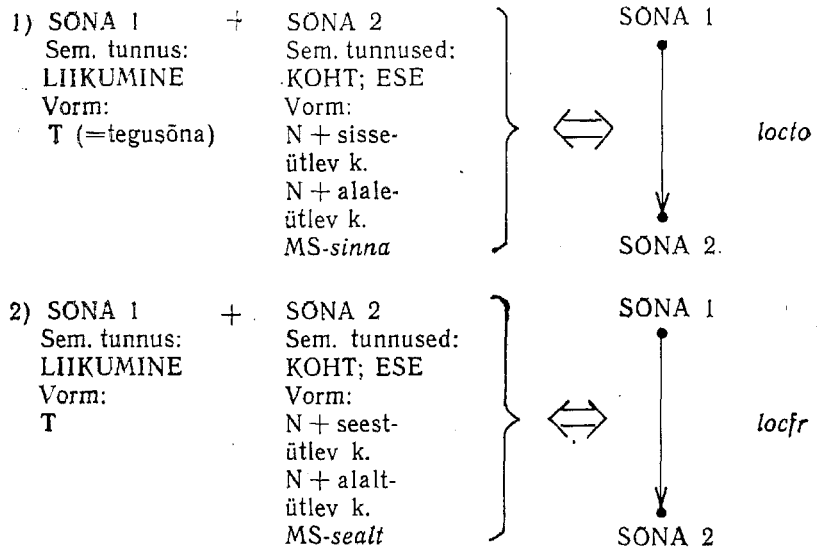
Esimest tüüpi reeglite ülesanne on leida lauses sõnapaarid, mille vahel saab kehtestada mingi süntaktilise seose. Filtrid on aga reeglid, mis kontrollivad sõltuvusmallide ja esimese reeglitüübi abil moodustatud struktuure nende vastavuse seisukohalt mitmesugustele antud keeles kehtivatele kitsendustele ning filtreerivad välja neid nõudeid mitterahuldavad struktuurid.

Süntaktilisi seoseid kehtestavate reeglite ühe osa moodustavad protseduurid, mis kõrvutavad lauses leiduvate sõnade sõnastikuartikleid, et kindlaks teha nende sobivust täitmaks üksteise sõltuvusmallide tühje valentse (vt. eespool toodud näidet sõnaga *viima 1*).

Teine osa süntaktilisi seoseid kehtestavatest reeglitest on sõnas-

[34] Eesti keele tegusõnade sõltuvusmallidest ülevaate saamiseks võib asjast huvitatuile soovitada raamatut H. Rätsep. Eesti keele lihtlausete tüübid. Tallinn, 1978. Selles leiduva materjali esitamisel pole küll silmas peetud masintõlget, kuid esituse põhimõtted on lähedased sellele, mida nõuab keele kirjeldus masintõlkesüsteemis. Raamatu põgusastki lehitsemisest piisab, et saada pilti eesti keele tegusõnade sõltuvusmallide süsteemi keerukusest.

tatud üldisemates terminites ning mõeldud eelkõige selliste seoste leidmiseks, mis sõnade sõltuvusmallides ei kajastu. Nende reeglite tüüpiline sõnastus on järgmine: kui kahest sõnast SÕNA 1 ja SÕNA 2 üks omab semantilisi tunnuseid A_1 ja vormilisi (süntaktilisi ja morfoloogilisi) tunnuseid B_1, kuna teine omab semantilisi tunnuseid A_2 ja vormilisi tunnuseid B_2, siis võib nende vahel kehtestada süntaktilise seose R. Näitena võib tuua kaks reeglit:



Neis reegleis tähistab MS-*sinna* kohamäärsõnu klassist «sinna», s. o. määrsõnu, mis väljendavad liikumise sihtkohta: *sinna, kaugele, kusagile, eemale,* MS-*sealt* aga määrsõnu nagu *sealt, kaugelt, kusagilt, eemalt.* Süntaktiline seos *locto* on seos tegevuse (liikumise) ja selle sihtkoha vahel, *locfr* aga seos liikumise ja selle lähtekoha vahel. Need reeglid konstateerivad niisiis, et iga liikumist väljendava tegusõna juurde võib kuuluda laiend, mis osutab liikumise lähtekohta (kust liigutakse), ja laiend, mis osutab liikumise sihtkohta (kuhu liigutakse). Erinevalt sõltuvusmallidest kehtivad need reeglid tervete sõnaklasside, seejuures eelkõige semantiliselt määratletud klasside kohta.

Filtrid on reeglid, millega kontrollitakse eelmist tüüpi reeglite abil konstrueeritud struktuuride vastavust mitmesugustele antud keeles kehtivatele kitsendustele. Need kitsendused võivad olla üldise iseloomuga ning rakenduda kõigile lausetele või suuremale osale lausetest, siis on tegemist globaalsete filtritega. Nad võivad aga olla seotud üksikute sõnarühmade või isegi

üksiksõnadega, üksikute positsioonidega lauses, siis räägitakse lokaalsetest filtritest. Esimest liiki kitsenduste näide on reeglid, mis reguleerivad sõnade järjekorda lauses. Eesti keeles on sõnade järjekord suhteliselt vaba, kuid on keeli (näiteks inglise keel), kus see on märksa rangemalt reeglistatud. Globaalsete filtrite hulka kuuluvad ka mitmesugused formaalse iseloomuga nõuded konstrueeritavate sõltuvusstruktuuride suhtes. Näiteks ei ole loomuliku keele lausetes üldreeglina võimalik, et üks sõna sõltuks samaaegselt kahest erinevast sõnast. See tähendab, et kui süntaktiliste seoste määramise käigus on saadud näiteks struktuur A→B←C, siis on alust arvata, et käesolev analüüsivariant on väär. Teine lausete sõltuvusesituste võrdlemisi universaalne formaalne omadus on nende projektiivsus. Piltlikult seletades seisneb omadus selles, et kui me lauses tõmbame süntaktiliselt seotud sõnade vahele nooled, siis need ei tohi omavahel ristuda; näiteks struktuur A B C D on mitteprojektiivne struktuur. Mitteprojektiivsed struktuurid on igas keeles lubatud vaid üksikutel kindlatel juhtudel, nii et kui lause analüüsi käigus kujuneb kusagil selline struktuur, siis enne käsiloleva analüüsivariandi jätkamist tuleb kontrollida, kas on tegemist lubatud või lubamatu projektiivsusega.

Lokaalse filtri näitena võib tuua juba eespool kirjeldatud kitsenduse sõna *käskima* laiendite vormivariantide omavahelise kombineerumise kohta. See on kitsendus, mis iseloomustab ainult antud sõna või väikest rühma sõnu.

Oleme linnulennult kirjeldanud masintõlkesüsteemi süntaksikomponendi keelevara. Püüame nüüd konkreetse näite varal anda mingi pildi sellest, kuidas erinevat liiki andmed koos töötavad. Vaatame näiteks, kuidas tehakse kindlaks lause *Isa käskis poisil jalgratta õue viia* süntaktiline struktuur. Morfoloogiakomponendis on selle lause sõnavormid varustatud järgmise informatsiooniga:

isa	(N; ainsus; nimetav k.)
	(N; ainsus; omastav k.)
	(N; ainsus; osastav k.)
käskima	(T; lihtminevik; 3. pööre)
poiss	(N; ainsus; alalütlev k.)
jalgratas	(N; ainsus; omastav k.)
õu	(N; ainsus; omastav k.
	(N; ainsus; osastav k.)
	(N; ainsus; sisseütlev k.)
viima	(T; DA-INF.)

Järgnevalt on vaja määrata sõnadevahelised süntaktilised seosed, võttes aluseks sõnade morfoloogilise vormi, sõnastikus leiduva informatsiooni ja süntaktilise analüüsi reeglid.

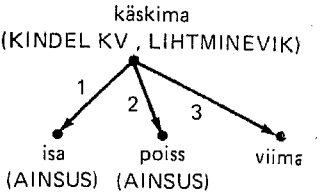
Esimene sõna, millel on sõnastikus oma sõltuvusmall, on sõna *isa:* see eeldab inimest tähistavat omastavas käändes nimisõna. Niisugust sõna aga lauses pole, seega jääb kõnealune aktandipositsioon täitmata. Järgmiseks sõnaks, millel on sõltuvusmall, on sõna *käskima*.

Sellel on tema sõltuvusmallis toodud kolm aktanti. Esimeseks neist peab olema inimest tähistav nimisõna nimetavas käändes. Ainus sõna lauses, mis seda nõuet rahuldab, on sõna *isa*. Selle sõna kolmest tõlgitsusest, millega ta on morfoloogiakomponendis varustatud, valitakse süntaksikomponendis seetõttu tõlgitsus «ainsus; nimetav kääne» ja kehtestatakse 1. aktandiseos *käskima* ja *isa* vahel.

Teise aktandina on sõna *käskima* sõltuvusmallis märgitud inimest või looma tähistav sõna, mis peab seejuures olema nimisõna osastavas või alalütlevas käändes. Neid nõudeid rahuldavad *isa* osastavas käändes ja *poiss* alalütlevas käändes. Kuivõrd *isa* on juba hõivatud, siis jääb üle vaid *poiss* alalütlevas käändes (osastava käände vormina on analüüsitav veel vorm *õue*, kuid see ei rahulda semantilisi tingimusi — tema semantiliseks kategooriaks sõnastikus on KOHT). Niisiis kehtestatakse teine aktandiseos *käskima* ja *poiss* vahel.

Kolmanda laiendina on sõna *käskima* sõltuvusmallis esitatud tegevust tähistav väljend, mis peab esinema *da*-tegevusnimena või *et*-kõrvallausena. Analüüsitavas lauses leidub üks *da*-tegevusnime vorm: *viia*. See rahuldab ka semantilisi tingimusi, omades semantilist tunnust TEGEVUS, ning järelikult sobib *käskima* kolmanda aktandi kandidaadiks. Niisiis on selleks ajaks konstrueeritud süntaktiline struktuur joonisel 9.

Konstrueeritavas süntaktilises struktuuris säilivad sõnadel vaid



Joon. 9

need morfoloogilised kategooriad, mida väljundkeele tõlkes on vaja arvestada.

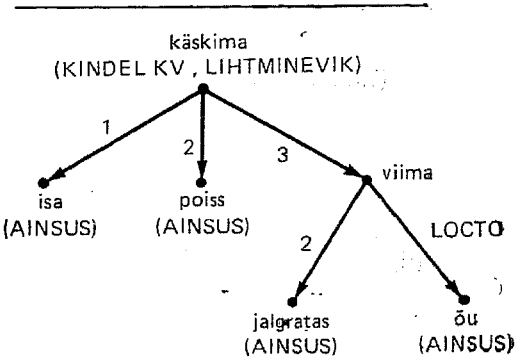
Sõna *käskima* sõltuvusmalli «tühje kohti» täitvate sõnade kandidaadid on leitud, niisiis otsitakse järgmine sõna, mille sõltuvusmall nõuab kindlas vormis ja kindlate semantiliste tunnustega laiendite olemasolu lauses. Niisugune sõna on *viima*.

Sõnal *viima* eespoolkirjeldatud tähenduses on sõnastikus kolm aktanti: kes viib, mida või mille viib ja millega viib. Esimeseks aktandiks sobib nimisõna nimetavas käändes semantilise tunnusega INIMENE või LOOM. Niisugust vaba sõnavormi lauses enam ei leidu. Nüüd tuleb kontrollida, kas ja mis tingimustel sõna *viima* esimese aktandi koht võib jääda täitmata. Ilmselt peab eesti keele kirjelduses leiduma reegel-filter, mis ütleb, et verbi *da*-tegevusnimelise vormi juures võib esimene aktant jääda osutamata. Leidnud sellise reegli, minnakse edasi ja asutakse otsima *viima* teise aktandi kandidaati — nimisõna ainsuse või mitmuse nimetavas või osastavas või ainsuse omastavas käändes, millel oleks

semantiline tunnus ESE, ELUSOLEND või AINE. Seni vabadest sõnadest rahuldab neid nõudeid sõna *jalgratas* omastavas käändes (*õu* puhul ei sobi semantiline tunnus). *Viima* viimase aktandi — vahendi — koht tuleb jällegi jätta täitmata, sest sobivat sõna lauses ei leidu.

Et aga lauses on veel vorm *õue*, mida sõnade sõltuvusmallide põhjal ei õnnestunud siduda ühegi teise sõnaga, tuleb pöörduda süntaktiliste seoste kehtestamise üldiste reeglite poole: tuleb otsida reegleid, kus üheks kirjeldavaks sõnaks on nimisõna ainsuse omastavas, osastavas või sisseütlevas käändes, evides seejuures semantilist tunnust KOHT, ning valida nende hulgast reegel või reeglid, mille teine sõnakirje sobib mõne analüüsitavas lauses leiduva sõna andmetega. Niisuguseid reegleid võib muidugi olla mitmeid ning siis tuleb jällegi pöörduda filtrite poole, et maksimaalselt kitsendada võimalike analüüsivariantide hulka. Antud juhul sobib näiteks eespool toodud reegel, mis kehtestab seose *locto* sõnade *viima* (kuivõrd selle semantiliste tunnuste hulka kuulub LIIKUMINE) ja *õu* vahel. Niisiis oleme lõpuks saanud analüüsitava lause süntaktiliseks esituseks struktuuri joonisel 10.

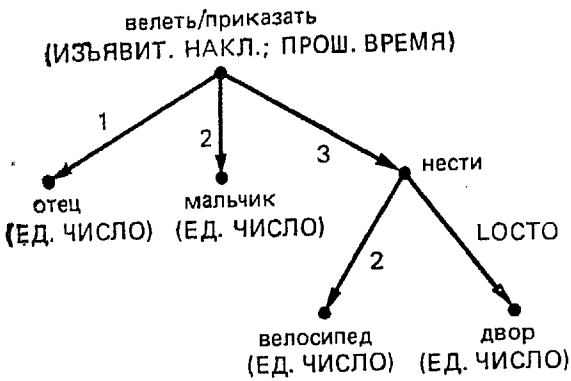
Analüüsiprotsessi kirjeldus oli muidugi äärmiselt lihtsustatud, kuid eesmärgiks oligi selgitada ainult analüüsi põhimõttelist kulgu. Tahaksime eriti juhtida tähelepanu asjaolule, et sõnavormide morfoloogilise struktuuri mitmesus laheneb süntaktilise analüüsi käigus



Joon. 10

peaaegu automaatselt, muidugi kui pole tegemist tõepoolest mitmetähendusliku lausega (nagu *Ema otsis isa*). Samuti väärib märkimist semantika arvestamise olulisus — tarvitseb vaid jälgida, mitmel korral sellegi lihtsa lause analüüsimisel läks vaja sõnade semantilisi tunnuseid, et valida õige analüüsivariant.

Kui on jõutud lähtekeele lause süntaktilise struktuurini, siis järgnebki masintõlkesüsteemis selle struktuuri t õ l k i m i n e vastavaks väljundkeele süntaktiliseks struktuuriks. Lausete süntaktilise struktuuri esituse eespool kirjeldatud tasand on juba peaaegu universaalne. Eriti on süntaktiliste seoste hulk valitud nimelt selle eesmärgiga, et see oleks kui mitte universaalne, siis vähemalt ühesugune lähte- ja väljundkeelele. Seetõttu seisneb üleminek ühelt keelelt teisele sellel süntaktilise struktuuri tasandil valdavalt struktuuri sõlmedes olevate sõnade asendamises



Joon. 11

väljundkeele sõnadega. Näiteks analüüsitud lause puhul ei tekitaks vist erilisi raskusi asendada eesti keele sõnad saadud struktuuris vene keele sõnadega: *isa — отец, käskima — приказать/велеть, poiss — мальчик, viima — нести, jalgratas — велосипед, õu — двор.* Tulemuseks on struktuur joonisel 11. Ning see on juba vene keele süntaksi- ja morfoloogiakomponendi vormistamise probleem, kuidas niisugust struktuuri venekeelse lausena väljendada.

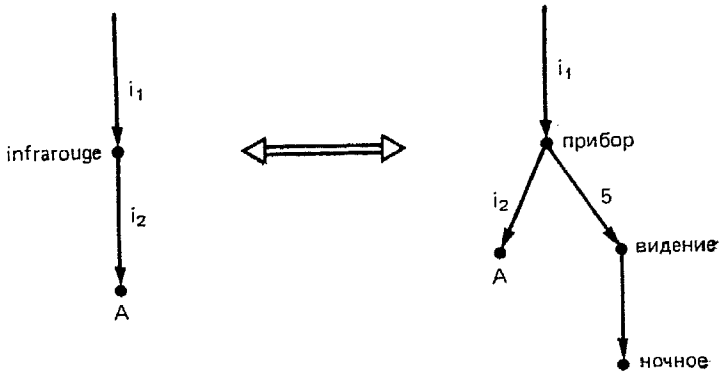
Niisugune sõnasõnaline asendamine ei ole siiski alati võimalik. Sageli on vaja erilisi reegleid üleminekuks ühelt keelelt teisele. Eespool viidatud prantsuse-vene tõlkesüsteemis on kasutusel kolm niisuguste reeglite tüüpi.

1) Ühe lähtekeele sõna asendamine väljundkeeles terve fraasiga, seega ka terve struktuurifragmendiga. Näiteks prantsuskeelsele terminile *infrarouge* vastab vene keeles väljend *прибор ночного видения.* Vastavalt sellele korralduvad tõlkimisel ümber ka süntaktilised seosed, nagu näha jooniselt 12.

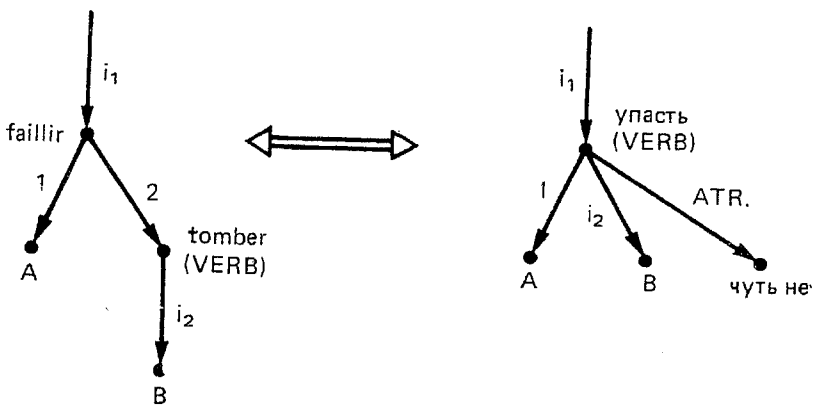
2) Lähtekeele struktuuri asendamine väljundkeele üheainsa üksusega. Siiagi sobib eelmine näide, ainult eeldusel, et lähtekeeleks on vene keel ja väljundkeeleks prantsuse keel.

3) Mitmesugused keerulisemat tüüpi ümberkorraldused, kus lähtekeele mingi struktuur teisendatakse väljundkeeles teise ehitusega struktuuriks. Näiteks prantsuse keele fraas *il faillit tomber* 'Ta pidi peaaegu kukkuma' antakse vene keeles edasi väljendiga *он чуть не упал,* millel on mitmes suhtes teistsugune struktuur kui prantsuskeelsel fraasil (joon. 13).

Eespool rõhutasime, et praegustes masintõlkesüsteemides püütakse keelevara vormistada võimalikult sõltumatuna konkreetsetest analüüsi- ja sünteesialgoritmidest.



Joon. 12



Joon. 13

See tähendab muu hulgas, et näiteks sõnade sõltuvusmallid, aga ka üldised süntaktilisi seoseid määratlevad reeglid on rakendatavad mitte ainult analüüsil, vaid ka sünteesil. Reeglid tuleb vaid ümber pöörata. See, mis analüüsis on reeglite sisendiks (morfoloogilised vormid), on sünteesi juhul väljundiks, kuna analüüsis väljundiks olevad süntaktilised konstruktsioonid (kindla süntaktilise seose abil seotud sõnapaarid) on sünteesialgoritmide lähtematerjaliks. Ei ole raske näha, kuidas eespool toodud sõnade nagu *käskima* ja *viima* kirjeldused või reeglid nagu (a) ja (b) on kasutatavad lausete sünteesimise käigus. Seepärast me ei hakkagi sünteesietapil põhjaliku-

malt peatuma. Tõhusate analüüsialgoritmide olemasolu korral on vastavad sünteesialgoritmid alati märksa vähem probleeme tekitanud.

3.3. Mõni sõna masintõlkest tänapäeva maailmas

Nagu eespool mainisime, ei ole seni veel jõutud masintõlkesüsteemideni, mis suudaksid tagada sajaprotsendiliselt korrektse tõlke. Mitmeid süsteeme kasutatakse siiski juba praegu piiratud ulatuses praktilise otstarbega tõlgete saamiseks. Näiteks tõlgitakse 1976. aastast peale Kanadas ilmateateid inglise keelest prantsuse keelde masintõlkesüsteemi abil. Süstemaatiliselt kasutatakse masintõlget ka USA Aatomienergia Komisjonis ja NASA-s (Rahvuslik Aeronautika- ja Kosmosevalitsus) tõlkimaks vene keelest inglise keelde. Masintõlget katsetati muide ka «Sojuz»—«Apollo» kosmoseeksperimendi ajal 1976. a.

Mitmete süsteemide loojad Läänes on kuulutanud, et nende süsteemid on valmis ulatuslikuks praktiliseks ekspluateerimiseks. Viimastel aastatel on kõige laiemalt reklaamitud USA-s loodud süsteemi SYSTRAN, mille praegune valdaja on kõlava nimega organisatsioon *World Translation Center, Inc.* asukohaga La Jollas Californias. Selle süsteemi väljatöötamise ja täiustamisega on tegeldud 1950-ndate aastate lõpust alates. 1970-ndail aastail on SYSTRAN-i lühema või pikema aja jooksul katseliselt kasutanud mitu asutust ja organisatsiooni, nende hulgas NASA ning USA õhujõud venekeelsete kosmosetehnika-alaste tekstide inglise keelde tõlkimiseks.

SYSTRAN on tema loojate iseloomustuse järgi kolmanda põlvkonna masintõlkesüsteem, mille programmvara on selles mõttes universaalne, et ei ole orienteeritud mingite konkreetsete keelte vahel tõlkimiseks. Teda saab kasutada tõlkimaks ükskõik mis keelest ja ükskõik mis keelde, mille kirjeldus on SYSTRAN-i programmide poolt nõutavas vormis süsteemile lisatud. Seepärast on võimalik SYSTRAN-i abil tõlkida samaaegselt mitmesse keelde. Küll aga on tarvis iga uue keele lisamisel välja töötada vastavad masinsõnastikud. Tõlgitud tekst võidakse kohe trükkida, kuid on ette nähtud ka võimalus, et tõlgitud laused ilmuvad eelnevalt kuvari ekraanile, kus vigade korral on võimalik neid parandada, ning alles seejärel trükitakse tekst. Kirjanduse andmeil on süsteem suuteline tõlkima ja tõlgitud teksti välja trükkima kiirusega kuni 300 000 sõna tunnis.

1976. aastal otsustas Euroopa Majandusühenduse Komisjon võtta SYSTRAN kasutusele Majandusühenduse dokumentide tõlkimisel. Inglise-prantsuse, prantsuse-inglise ja inglise-itaalia versioonid on valmis ja leiavad juba kasutamist, käsil on inglise-saksa ja prantsuse-saksa moodulid.

Euroopas loodud süsteemidest on tuntumaid ja põhjalikumaid Prantsusmaal Grenoble'i ülikoolis arendatav süsteem venekeelse teadus- ja tehnikaalase kirjanduse tõlkimiseks prantsuse keelde. Selle esialgset varianti katsetati juba 1960-ndate aastate lõpus, käesoleval ajal täiustatakse süsteemi keelevara, eriti semantilist analüüsi.

Saksa FV-s on masintõlkesüsteemide väljatöötamine käsil mitmes ülikoolis — Saarimaa ülikoolis Saarbrückenis, Heidelbergi ülikoolis ning Ruhri ülikoolis Bochumis.

1974. a. asutati rahvusvaheline töögrupp nimega «Leibniz», mis ühendab Prantsusmaa, Saksa FV, Itaalia ja Šveitsi masintõlkega tegelevaid kollektiive. Töögrupi üks eesmärke on välja töötada ühtne metoodika ja formaalsete vahendite süsteem, mis oleks rakendatav erinevate keelte kirjeldamisel, nii et kõikide nende keelte vahel tõlkimiseks saaks kasutada samu analüüsi- ja sünteesialgoritme.

Aasias tegeldakse masintõlkega tõsisemalt ilmselt vaid Jaapanis, Kioto ja Kiušiu ülikoolides. Kioto ülikooli süsteemi omapära seisneb selles, et tõlgitud teksti võib väljastada ka sünteesitud kõnena. Veel töötab Hongkongi ülikoolis masintõlkesüsteem, mis tõlgib hiinakeelseid matemaatika- ja füüsikaalaseid ajakirju inglise keelde.

Nõukogude Liidus tehtud masintõlkealasest tööst saime pildi juba peatüki alguses. Üleliidulises Teadus- ja Tehnikakirjanduse ning Dokumentatsiooni Tõlkekeskuses on välja töötatud praktiliseks ekspluateerimiseks mõeldud inglise-vene tõlkesüsteem АМПАР-2, samuti prantsuse-vene tõlkesüsteem, mille aluseks on NSVL TA Rakendusmatemaatika Instituudis valminud eksperimentaalne süsteem (eespool toodud loendis kümnes).

Põhjalik ja hoolikalt läbimõeldud keelevara on prantsuse-vene tõlkesüsteemil elektrotehnikaalaste tekstide tõlkimiseks, millele viitasime eelmises jaotises. Töid juhib lingvisti-teoreetikuna, eelkõige semantikuna hästi tuntud J. Apresjan. Süsteemi vene keele sõnastik sisaldab üle 5300 tüve, mis on aluseks 130 000 sõnavormile. Prantsuse keele sõnastikus on 3500 sõnatüve, nende alusel moodustatakse üle 50 000 sõnavormi. Süsteemi algvariant ЭТАП-1 võeti kasutusele 1981. aastal. Samal ajal alustati töid uue variandi ЭТАП-2 kallal, milles täiustatakse süntaktilise analüüsi ja keeltevahelise vastavuse reegleid, samuti tööd mitmetähenduslike sõnade ja konstruktsioonidega. Esimese variandi tõlke kvaliteet oli hea, kiirus aga ebapiisav (lause analüüsiks kulus poolest kahe minutini). Uues variandis lubab keeleandmete otstarbekam jagamine analüüsireeglite ja sõnaartiklite vahel analüüsi kiirust tunduvalt tõsta.

Veel võib mainida jaapani-vene tõlkesüsteemi, mis on loomisel Moskvas NSVL TA Orientalistika Instituudis.

Loodud on ka kitsama temaatikaga tekstide tõlkimiseks mõeldud süsteeme, näiteks ingliskeelseid

naftakeemia-alaseid tekste tõlkiv СИМПАР, mis töötab Üleliidulises Naftakeemia Teadusliku Uurimise Instituudis alates 1981. aastast.

Masintõlkesüsteemide poolt tehtavate tõlgete kvaliteedi suhtes on raske teha mingeid üldistusi, on ju vastavad näitajad eri süsteemidel erinevad. Pealegi on seda küsimust vaid väheste süsteemide puhul spetsiaalselt uuritud. Üks neist on äsja mainitud prantsuse-vene masintõlkesüsteem Rakendusmatemaatika Instituudis. Et seda peetakse tugevaks teise põlvkonna süsteemiks, siis võib tema näitajaid lugeda küllalt tüüpilisteks.

Süsteemi poolt tõlgitud matemaatilised tekstid anti hindamiseks nii lingvistidele kui ka matemaatikutele. Tõlgitud lauseid lasti hinnata grammatilise õigsuse, arusaadavuse ja adekvaatsuse (originaalile vastavuse) seisukohalt. Hinnangud anti kolmepallises süsteemis — «hea», «keskmine» ja «halb». Umbes tuhat hinnatud lauset jagunesid järgmiselt. Arusaadavuse seisukohalt sai 63,5% lauseid hinde «hea», 25% olid keskmised ja 11,5% halvad, grammatilise õigsuse seisukohalt 49,2% head, 32,2% keskmised ja 18,5% halvad, adekvaatsuse seisukohalt 60,7% head, 25,6% keskmised ja 13,7% halvad.

Niisiis keskeltläbi poolte lausete tõlkimisel esinevad ühe- või teistsugused vead. Niisuguse süsteemi praktilisel kasutamisel tuleb seetõttu tõlgitud tekstid anda inimestele järelredigeerimiseks. Enamus süsteeme töötabki praegu just niimoodi. Masintõlke alal tehtava töö intensiivsus ning agarad püüded loomuliku keele semantiliste mudelite lülitamiseks süsteemidesse lubavad oletada, et juba 1980-ndate aastate keskel on masintõlge juba märksa tavalisem nähtus.

Lõpetuseks paar arvuti tõlkenäidet. Esimene näide on Grenoble'i süsteemi tõlge vene keelest prantsuse keelde, teine on Zürichis töötava süsteemi ADS poolt tehtud majandusalase teksti fragmendi tõlge saksa keelest inglise keelde.

1. Аналогично случаю молекул для нахождения функции нулевого приближения необходимо сначала решить задачу о рассеянии налетающей частицы неподвижными нуклеонами дейтрона.

Il est, par analogie avec un cas des molécules, nécessaire d'abord de resoudre le problème de la diffusion de la particule incident par les nucléons immobiles du deutéron pour que la fonction de l'approximation nulle soit trouvée.

2. Der Wettbewerbsbericht ist mittlerweile ein bedeutender und regelmässiger Bestandteil des Dialogs zwischen dem Europäischen Parlament und der Kommission geworden. Der vorliegende Bericht zieht eine Bilanz der Wettbewerbspolitik in diesem Jahr.

Meanwhile, the Report on Competition Policy has become an important and regular part of the dialogue between the European Parliament and the Commission. The present report draws a balance sheet of the competition policy in this year.

4. Keelt mõistvad süsteemid

Arvuti: Maffiasse kuuluvad kõige hullemad gangsterid.
Inimene: Teil on olnud tegemist maffiaga?
Arvuti: Nad spioneerivad minu järel.
Inimene: Miks?
Arvuti: Nad tahavad minu vastu valetõendeid fabritseerida.
Tegelikkuses asetleidnud dialoogist.

Tuntud inglise matemaatik, küberneetika üks rajajaid A. M. Turing pakkus omal ajal tehissüsteemil (arvutil) intellekti olemasolu kindlakstegemiseks testi, mille sisu seisneb selles, et inimene, vesteldes loomulikus keeles (näiteks teletaibi vahendusel) aruka süsteemiga, ei suuda kindlaks teha, et ta vestluskaaslane ei ole mitte inimene, vaid tehissüsteem. Tänu juhusele olevat vähemalt üks süsteem niisuguse katse juba läbi teinud, tõsi küll, asjaoludel, mis olid tema jaoks võrdlemisi soodsad. See oli ameeriklase J. Weizenbaumi loodud «vestlev süsteem» ELIZA.

Üks äriagent soovis demonstreerida mõnele oma kliendile ajajaotusarvuti tööd. Ta leidis vaba terminali, kuid selle juures oli sedel «reserveeritud professor Bobrow'le». Agent üritas seejärel teletaibi kaudu professor Bobrow' poole pöörduda, kuid tegelikult sattus ühendusse süsteemiga ELIZA. Arenes järgmine dialoog:

Agent: Mis te arvate, kas ma võiksin täna hommikul arvutit kasutada?
Vastus: Milleks te seda küsite?
Agent: Võib-olla õnnestub mul teha mõned täiendavad müügitehingud.
Vastus: Kuid miks te selles kindel ei ole?
Agent: Mu kliendid pole veel kordagi näinud süsteemi demonstratsiooni.
Vastus: Kas see tähendab teile väga palju?
Agent: Muidugi tähendab.
Vastus: Te näite asjast väga huvitatud olevat.
Agent: See on naeruväärne!
Vastus: Mis te mõtlete sellega, et see on naeruväärne?
Agent: Kas on mingi põhjus, miks ma ei saa praegu arvutit kasutada?
Vastus: Rääkige mulle lähemalt oma klientidest.
Agent: Palun helistage mulle numbril 491-1950.[35]

Nagu näha, kulus agendil tükk aega, enne kui talle «hakkas koitma», et midagi on korrast ära. Te-

[35] Palme J. Making Computers Understand Natural Language. — In: Artificial Intelligence and Heuristic Programming / Ed. by N. Findler and B. Meltzer. Edinburgh, 1971, p. 200.

gelikult on ELIZA üks lihtsamaid semantilisi süsteeme, mis keele sisulisest mõistmisest on veel kaugel. Selle loomisel oli põhiline eesmärk imiteerida elektronarvutil inimestevahelist vestlust suvalisel teemal. Vestluse näiv ladusus saavutatakse võrdlemisi lihtsate vahenditega, mis on orienteeritud eelkõige vestluse välistele tunnustele. Peamine pole kaasvestleja lausete täpne mõistmine, vaid kindla tähendusega väljendite (võtmesõnade) ja kindlate lausetüüpide äratundmine. Iga võtmesõnaga ja lausetüübiga on süsteemi mälus seotud vastuste mudelid, millest lähtudes konstrueeritakse antud konteksti sobiv konkreetne vastus. Võib öelda (seda on öelnud ka süsteemi looja J. Weizenbaum ise), et süsteemi eesmärgiks pole miite niivõrd kaasvestleja mõistmine, kuivõrd mittemõistmise oskuslik varjamine. Tõepoolest, kui jälgida tähelepanelikumalt toodud dialoogi, siis võib veenduda, et ELIZA pöördumised on üles ehitatud väga üldiste suhtlemistrafarettide järgi. Üheski neist ei kajastu eelnenud pöördumiste otsene sisu, ELIZA laused haakuvad nendega niisuguste sisutute väljendite kaudu nagu *see* või *asi* *(miks te selles kindel ei ole, te näite asjast väga huvitatud olevat)*.

Niisugusena pole ELIZA siiski keelt mõistvate süsteemide tüüpiline esindaja. Nende süsteemide loomise põhieesmärk on olnud ja on ka praegu saavutada loomuliku keele tekstide sisuline mõistmine elektronarvutite poolt. Niisuguse inimliku omadusega varustatud süsteemide loomise taga on kaks suurt tõukejõudu.

Esimene on seesama praktikast lähtuv vajadus infotöötluse automatiseerimise järele, mis kõige selgemini ja vahetumalt avaldub automatiseeritud infootsisüsteemide ning — pisut kaudsemalt — ka masintõlkesüsteemide loomises. Sellelt seisukohalt vaadelduna on keelt mõistvad süsteemid niisugused süsteemid, mis suudavad esitatud küsimustest sisuliselt aru saada ja nendele ka sisuliselt vastata. Nad on seega faktiotsisüsteemide täiuslikum tulevikuvariant.

Keelt mõistvate süsteemide praktilisel vajalikkusel on veel teine aspekt, millest samuti oli üksikasjalikult juttu esimeses peatükis: vajadus kasutada loomulikku keelt arvutiga suhtlemiseks mitmesuguste ülesannete lahendamisel. Osutasime, et küllalt suure ülesanneteringi puhul ei oleks loomulik keel mitte ainult kõige mugavam, vaid ka kõige otstarbekam suhtlemisvahend inimese ja arvuti vahel. Sel juhul on keelt mõistev süsteem vajalik kui vahendajasüsteem inimese ja vastava ülesandeid lahendava automatiseeritud süsteemi vahel — kui süsteem, mis «tõlgib» inimese poolt formuleeritud ülesanded arvutile mõistetavasse programmkeelde.

Juba viimati käsitletud praktiline ülesanne tõstatab küsimuse sellest, milles õieti seisneb keele mõistmine ning missugusel määral on võimalik saavutada arvutilt mõistmist samas mõttes, nagu me sellest räägime inimestevahelise suhtlemise puhul. See on ka oma-

ette vaadatuna äärmiselt oluline ja huvitav probleem.

Siin peitubki teine tõukejõud keelt mõistvate süsteemide loomiseks — huvi keele mõistmise kui teoreetilise probleemi vastu. Loomulik keel on ainult inimesele omane nähtus, ning võimet kasutada niisugust keelt suhtlemiseks, seda mõista, loetakse inimintellekti üheks iseloomulikuks omaduseks. Katsed modelleerida keele mõistmist on osa püüdest realiseerida arvutil niisuguseid intellektuaalseid funktsioone, mis on seni olnud omased vaid inimesele. Viimane püüd on tänapäevaks kujundanud kui mitte iseseisva teadusharu, siis omaette uurimisala, küllalt kindlapiirilise, et talle on jõudnud kinnistuda oma nimetus — t e h i s i n t e l l e k t.[36] Keele mõistmise modelleerimine arvutil keelt mõistvate süsteemide loomise näol on niisiis üks tehisintellekti alla kuuluvaid probleeme. Enne keelt mõistvate süsteemide konkreetsete küsimuste juurde asumist on seetõttu kasulik lühidalt selgitada, mida see laiem kontekst haarab ja mida ta tähendab keele automaatse analüüsi ja sünteesi meetodite väljatöötamisele.

4.1. Keele mõistmine ja tehisintellekt

Inimesi on alati köitnud mõte võõra, mitteinimliku mõistuse olemasolu võimalikkusest. See idee kordub inimkonna loomingus ikka ja jälle alates ürgvanadest legendidest, kus räägitakse üleloomulike võimetega olenditest, kuni tänapäeva ulmekirjanduseni või kosmosetsivilisatsioonide otsinguteni teaduslikel alustel.

Teiselt poolt väidetakse, et inimese suure tehnikahuvi, masinate ja muude tehisasjade loomise taga on olnud püüd **pääseda** loodusest, mida ta ei suuda lõpuni mõista, ning ümbritseda end masinate ja asjadega, mis ta ise on loonud ja mida seetõttu mõistab ning suudab valitseda. Inimkonna ajalugu tunneb palju legende, aga ka fakte katsetest luua masinaid, mis suudaksid arukalt tegutseda, seega omaksid mõistust, kuid samal ajal kuuletuksid täielikult oma loojatele.

Üks tuntud legende väidab näiteks, et XV sajandil olevat Praha rabi Löw savist valmistanud ja hingestanud inimesetaolise olendi, Golemi. Golem ei olnud võimeline

---

[36] Termin *tehisintellekt* uurimisala tähistajana on kahtlemata omapärane. Tähistab termin ju algselt seda, mida kõnealuste uurimuste tulemusena püütakse luua, s. o. (inim)intellekti teatud aspektide tehismudelit. Uurimisala — tehisintellektiteaduse — tähistamiseks oleks loogiliselt vaja võtta kasutusele teine termin. Kuid loogikast hoolimata on *tehisintellekt* kujunenud just nimelt uurimisala tähistajaks (inglise keeles *artificial intelligence*, vene keeles *искусственный интеллект*). Üks põhjusi on ilmselt puht keeleline: kõnealustest väljenditest on kõigis keeltes raske mingil küllalt levinud viisil (eesti keeles näiteks liidete *-loogia* või *-ika* abil) moodustada adekvaatset terminit. Kõne alla võiks siiski tulla *intellektika;* selle kenasti kõlava termini ainus puudus on vahest selles, et eesliite *tehis-* sisu pole leidnud otsest väljendust.

midagi tegema omal algatusel, kuid täitis kõik oma isanda käsud. Nagu naljatades märgib B. Raphael raamatus «Mõtlev kompuuter», et mitmed tänapäeva elektronarvutite loomise ja kasutamise ajaloost tuntud mehed, nagu J. von Neumann, N. Wiener ja M. Minsky (tehisintellektialaste uuringute üks alusepanijaid), olevat kuulsa Praha rabi järeltulijad.

Golem on muidugi legend. Kuid fakt on see, et inimest on alati huvitanud mõistuse ja mõtlemise saladused. Fakt on ka see, et sellest ajast saadik, kui arvutite võimed said laiemalt tuntuks, on vaieldud selle üle, kas nende abil saab imiteerida inimese mõtlemisvõimet, või siis luua hoopis teist tüüpi mõistust, «masinmõistust». Sellest ajast pärineb ka termin *tehisintellekt.*

Tänapäeval ei diskuteerita enam nii väga selle üle, kas mõtlevate masinate loomine on põhimõtteliselt võimalik, samuti ka mitte selle üle, missuguseid ohte inimkonnale kätkeks niisuguste masinate ilmumine. Seda mitte niivõrd sellepärast, nagu oleks küsimus nüüd üheselt selge. See ei saagi nii olla, sest ühelt poolt pole selge, kuidas täpselt määratleda mõtlemise olemust, teiselt poolt aga pole kaugeltki selge ka see, mida on võimalik ja mida ei ole võimalik teha arvutite abil. Põhjus on pigem asjaolus, et tehisintellekti näol on välja kujunenud küllalt kindel ja konkreetne probleemide ja ülesannete ring, mis puudutab intellektuaalse tegevuse mingeid tahke. Kõigi nende puhul on selge, et m i n g i l määral on vastavate intellektuaalsete funktsioonide üleandmine arvutile võimalik. M i s s u g u s e l määral see üleandmine intellekti ühe või teise avalduse puhul lõpuks õnnestub, sõltub eelkõige meetoditest, mida uurijad on suutelised välja töötama.[37]

Tehisintellekti alla kuuluvaid uuringuid on liigitatud mitmeti. Tegelikult polegi need ühest kindlast alusest lähtuvad liigitused, vaid probleemide loetelu, millega faktiliselt tegeldakse ja mille puhul on jõutud mingite meetodite väljatöötamiseni.[38]

Esimestena said ilmselt tuntuks «mängivad» arvutid — süsteemid, mis olid programmeeritud nii, et suutsid kõrvalise abita mängida kabet, malet või teisi mänge. Tänapäeval on tuntud ka süsteemid, mis võivad luua muusikat (valsse, so-

---

[37] Jätkuvad muidugi ka filosoofilised diskussioonid, kuid needki on laskunud märksa konkreetsemale tasemele. Vastavate huvidega lugejale soovitame tutvuda diskussiooniga teemal «Inimene-masin-süsteemide sotsiaalfilosoofilised probleemid: tehisintellekt kui kompleksne teaduslik-tehniline probleem», mille korraldas ajakiri «Вопросы философии». Diskussiooni materjalid on avaldatud ajakirja 1979. a. 2., 3. ja 4. numbris.

[38] Meil pole võimalik anda tehisintellektist mingit süstemaatilist ülevaadet. Eesti keeles sobiv ülevaateteos puudub, kuid vene keeles on ilmunud mitmeid häid raamatuid. Eriti võib soovitada eespool mainitud teost Б. Рафаэл. Думающий компьютер. М., 1979. 407 с. Selle inglise keelest tõlgitud raamatu autor B. Raphael on Stanfordi uurimisinstituudi tehisintellektilabori, maailma ühe tuntuma tehisintellektikeskuse direktor.

naate, fuugasid vms.), joonistada pilte, esineda geograafia- või keeleõpetajatena või projekteerijatena, diagnoosida haigusi, programmeerida, s. o. koostada elektronarvutile programme teatud tüüpi ülesannete lahendamiseks. Ja muidugi on tuntud «autonoomsed intelligentsed süsteemid» — robotid.

Tehisintellektisüsteemide loetletud tegevused võivad iseenesest olla võrdlemisi mittevajalikud (milleks on näiteks vajalik kabet mängiv või valsse komponeeriv kompuuter?), võivad aga olla ka väga suure praktilise tähtsusega, nagu seda on näiteks haiguste diagnoosimine või programmeerimine. Tehisintellekti enda seisukohalt pole aga olulised välised tegevused, vaid tegutsemispõhimõtted, meetodid, mis on tegevuste aluseks. Kõigepealt tuleb kindlaks teha vastavate meetodite põhimõtteline olemus. Seejärel on vaja need meetodid formaliseerida, selleks et nad oleksid arvutil realiseeritavad. Just siin peitub tehisintellektisüsteemide ees seisvate ülesannete keerukus, ühtlasi see, mis ühendab niivõrd erinevatele tegevusaladele orienteeritud süsteeme. Formaliseerimisel selgub, et samu meetodeid, mida kasutatakse kabe- või malemängus, saab suurema või väiksema eduga rakendada projekteerimisel, haiguste diagnoosimisel, programmeerimisel või ka loomuliku keele tekstide mõistmisel.

Käsitledes küsimust tehisintellekti uurimisobjektist — sellest, m i d a tegelikult on vaja modelleerida, — arutleb näiteks D. Pospelov, tehisintellektialaste uuringute üks algatajaid Nõukogude Liidus, järgmiselt. Kui ma olen kirjutanud väga hea malet mängiva programmi ja sisestanud selle arvutisse, kas siis võib öelda, et arvuti, mis mängib selle programmi järgi malet, on näide tehisintellektist? Ilmselt mitte. Kui sellele programmile lisada näiteks muusikat komponeeriv programm, nii et arvuti mitte ainult ei mängi hästi malet, vaid komponeerib ka päris häid valsse ja marsse, kas on siis rohkem põhjust nimetada teda tehisintellektiks? Ilmselt samuti mitte. Pole vaja modelleerida mitte üksikuid väliseid intellekti funktsioone omaette — malemängu, muusika komponeerimist jm. —, vaid «globaalseid psühholoogilisi mehhanisme, mis võimaldavad k o n s t r u e e r i d a neid protsesse».[39]

Need «globaalsed psühholoogilised mehhanismid», mis on tehisintellektialaste uuringute võtmeprobleemideks, ükskõik missugustele rakendustele need on suunatud, võib tinglikult jagada kolme suurde rühma. Esimesse kuuluvad probleemid, mis on seotud mälu ja teadmistega: igal intelligentsel süsteemil peab olema ulatuslikke teadmisi nähtustest, mis moodustavad tema «maailma» ning mida puudutavaid ülesandeid süsteem lahendab. Teise rühma moodustavad nende teadmistega opereerimise meetodid: vajalike tulemusteni viivate deduktiivsete ja mittededuk-

[39] П о с п е л о в Д. А. Профессионально и проблемно ориентированные интеллектуальные системы. — Вопросы философии, 1979, № 2, с. 58.

tiivsete arutluskäikude aluseks olevad põhimõtted. Kolmandaks peab intelligentsel süsteemil olema «eesmärkide plokk», mis juhib nii teadmiste talletamist ja nende korraldamist mälus kui ka nende kasutamist arutluskäikudes, samuti arutluskäikude endi korraldust ülesannete lahendamisel.

Olles põgusalt iseloomustanud seda, mida taotleb tehisintellekt, võime nüüd tagasi pöörduda keelt mõistvate süsteemide juurde ja küsida, mis rolli etendab keele mõistmise modelleerimine tehisintellektis tervikuna. Kui ütleme, et keel ja tema mõistmise protsessid on tehisintellektis viimasel ajal omandanud erilise koha, siis ei tule see sugugi meie soovist oma uurimisala kuidagi esile tõsta (ehkki see soov oleks loomulik).

D. Pospelovi eespool tsiteeritud mõtteavaldusele järgneb originaaltekstis vahetult järgmine lõik: «Sellest [s. o. vajadusest modelleerida eespool mainitud globaalseid psühholoogilisi mehhanisme — *H. Õ.*] ongi tingitud tähelepanu, mida tehisintellektisüsteemide loomise spetsialistid osutavad keelt mõistvatele süsteemidele, mis on suutelised suhtlema inimesega viimasele harjumuspärases keeles. Loomulik keel on kõige võimsam meile seni teadaolevatest tegelikkust modelleerivatest süsteemidest. Siit ka loomulik soov kasutada seda süsteemi tulevastes «intelligentsetes» robotites. Sel juhul on kõik, mida loomuliku intellekti kandjad suudavad väljendada sõnadega, viidav ka tehissüsteemi.»

Loomulik keel on inimese tunnetus- ja mõtlemisprotsesside vahetumaid avaldusi. Kõige selle kajastamiseks — modelleerimiseks —, mis saab olla nende protsesside sisuks ning mida ühtlasi on vaja väljendada suhtlemises teiste inimestega, peavad leiduma keeles omad vahendid. Just selles mõttes on keel «kõige võimsam modelleeriv süsteem». Igale teooriale, mis püüab seletada inimese mälu, tunnetusprotsesside või loogiliste arutluste struktuuri, pakub keel rikkalikku ja seejuures täiesti objektiivset ning kontrollitavat materjali.

On võrdlemisi ulatuslik klass ülesandeid, mida tehisintellektisüsteemides püütakse lahendada ja mida ühendab asjaolu, et kõigi nende puhul on keskseks mõistmise probleem — mõistmine samas mõttes nagu keele puhul. Lugedes üksteisele järgnevaid sõnu lauses ja üksteisele järgnevaid lauseid tekstis, peame mõistma, mida see tekst kui tervik tähendab: me peame interpreteerima teksti teatud viisil. Väliselt küll väga erinev, kuid sisult üpris lähedane on näiteks protsess, mille alusel arst määrab haigele diagnoosi: talle on antud teatud hulk objektiivseid näitajaid haige organismi talitluste kohta ning ta peab otsustama, mida need näitajad tähendavad — ta peab neid tõlgendama ja mõistma, missuguse haigusega on tegemist. Veelgi ilmsem on teksti mõistmise sarnasus kujundite visuaalse äratundmisega; ka selle protsessi automatiseerimine on tehisintellekti üks raskemini saavutatavaid ees-

märke. Meie nägemisanalüsaator jaotab nagu tehisanalüsaatorgi nähtava pildi suureks hulgaks komponentideks: joonteks, nurkadeks, kontuurideks, erineva heleduse-tumedusega piirkondadeks. Nendest komponentidest paneme seejärel kokku tervikpildi ning tunneme ära vastava objekti või situatsiooni: laua, tooli, maja, tuttava inimese näo. See tähendab, me interpreteerime seda visuaalsete komponentide kombinatsiooni ja mõistame teda kui tervikut, kui seda või teist eset. Nägemistaju automatiseerimisel, nii nagu ka näiteks automaatdiagnoosimisel, on osutunud kõige raskemaks just osadest terviku tuletamine, tõlgendamine, mõistmine, sest see eeldab süsteemilt ulatuslikke teadmisi ning arutlusvõimet.

Ka keeletekstide analüüsimisel on üks raskemaid probleeme, kuidas mõista teatud lausete järjendit kui sisulist tervikut, kui mingi tervikliku situatsiooni või sündmuse kirjeldust. Kuid keele puhul oleme siiski mitmes suhtes paremas olukorras kui näiteks visuaalsete piltide mõistmise automatiseerimisel. Keeleteksti puhul võime olla enam-vähem kindlad, et kõik, mis tekstis leidub, on kuidagi seotud selle tähendusega. Objektiivset pilti, eset või situatsiooni vaadeldes ei saa me kaugeltki ette teada, missugused selle osad etendavad ja missugused ei etenda mingit rolli pildi või situatsiooni mõtestamisel. Keele puhul on meil ette antud vähemalt teatud analüüsiüksused — sõnad ja laused. Ja keele puhul teame ka küllalt suurt hulka reegleid, mille järgi nende üksustega teksti mõistmisel opereeritakse. Nägemistaju käsitlemisel ei ole iseenesest sugugi selge, milliste elementidena inimene tajutavat kujundit analüüsib. Siin avaldub veel üks asjaolu, miks keel pakub tehisintellektisüsteemide loojatele erilist huvi: keele puhul on inimeses kulgevad psühholoogilised protsessid välja, nähtavale toodud, «objektiveeritud». Keeletekst on neid protsesse esitav objektiivne materjal.

Need ongi kaks peamist põhjust, miks keele mõistmine ja keelt mõistvad süsteemid tehisintellektispetsialistidele erilist huvi pakuvad: esiteks keele universaalsus inimese tunnetus- ja mõtlemisprotsesside kajastajana, teiseks asjaolu, et keeletekstid on vahetult antud objektiivne materjal, millest alustades ja mida otsekui suunava teemärgistusena kasutades on lihtsam jõuda nende psühholoogiliste mehhanismideni, mis tehisintellekti tegelikult huvitavad.

See kõik on nii, kui vaatame keelt mõistvaid süsteeme tehisintellekti laiemate huvide ja taotluste kontekstis. Süsteemide endi väljatöötamise seisukohalt vaadatuna on keele esimesena mainitud omadus, tema universaalsus, aga peamiste raskuste allikaks. Keele mõistmisel ei ole tegemist ühe või paari spetsiifilise infotöötluse vormiga nagu näiteks nägemistaju puhul. Keeletekstide mõistmises on kompleksselt esindatud kõik intellektuaalse infotöötluse vormid, alustades lihtsast mõistete identifitseerimisest (äratundmisest) ning lõpetades de-

duktiivsete ja induktiivsete arutluskäikude ning üldistustega. Neist raskustest aga räägime juba üheskoos keelt mõistvate süsteemide loomise muude probleemide käsitluga.

4.2. Esimesed süsteemid ja põhimõttelised probleemid

Keelt mõistvate süsteemide ülevaates nimetatakse enamasti kõigepealt süsteemi BASEBALL, mille 1960. aastal töötasid välja B. Green ja ta kaastöölised. See süsteem vastas lihtsustatud inglise keeles esitatud küsimustele, mis puudutasid andmeid USA pesapallimeeskondade omavaheliste mängude kohta 1959. aastal. Küsimused võisid olla võrdlemisi vabalt varieeritud, näiteks: *Kus mängis «Red Sox» juulikuus?, Missugused meeskonnad mängisid juunikuul Chicagos?, Mitu mängu võitsid meeskonnad oma koduväljakutel?*

Andmed BASEBALL-i mälus olid esitatud kindlaformaadiliste hierarhiliste struktuuridena, kus elementideks olid paarid «tunnus — väärtus», näiteks *meeskond = Red Sox, kuu = juuli, koht = Chicago* jne. Hierarhia kõrgeim tasand oli aasta, selle sees olid andmed korrastatud kuude järgi, iga kuu ulatuses kohtade järgi, kus sel kuul mängud toimusid, iga koha all aga olid esitatud konkreetsed andmed seal toimunud mängude kohta. Vastuse otsimine küsimusele seisnes andmete hierarhias sellise trajektoori leidmises, milles tunnuste—väärtuste paarid ühtisid küsimuses sisalduvate andmetega ning mis haaras ka seda tunnust, mille väärtust parajasti küsiti.

Süsteemi lingvistiline komponent — lingvistiline protsessor, nagu praegu öeldakse — teisendas iga küsimuse kujule, mis vastas andmete esitusviisile mälus. Keeleline analüüs koosnes küsimuse süntaktilise ja semantilise analüüsi etappidest. Kuid BASEBALLI-i keeleanalüüsi mehhanismid pakuvad praegu vähe huvi, seepärast pole põhjust neil üksikasjalikumalt peatuda. Süsteemi eesmärk oli eelkõige illustreerida tolle aja kohta uudsete andmete esitus- ja töötlusmeetodite rakendamise võimalusi arvutite mittestandardsetel kasutusaladel. Plaanitseti küll rakendada BASEBALL-i mitmesuguste hääletuste tulemusi puudutavatele küsimustele vastuste saamiseks, aga see projekt jäi realiseerimata.

Märksa suuremat huvi keeleväljendite sisu kirjeldusmeetodite seisukohalt pakub samast perioodist pärinev R. Lindsay loodud süsteem SAD-SAM. R. Lindsay on huvitatud ka keele mõistmise probleemi teoreetilistest aspektidest. Tema seisukohad neis küsimustes väärivad praegugi lähemat kirjeldamist eriti seetõttu, et nende najal selguvad mitmed probleemid, mis on eriomased keelt mõistvatele süsteemidele, erinevalt näiteks infootsisüsteemidest või masintõlkesüsteemidest. Vastandades oma lähenemist meetoditele, mida kasutatakse infootsisüsteemides ja ka näiteks BASE-

BALL-is, rõhutab R. Lindsay, et keele mõistmist ei saa taandada lauses leiduvate sõnadega kui märksõnadega opereerimisele: «Inimesele on omane intellektuaalne võime, mida võib iseloomustada sõnadega: ta m õ i s t a b selle m õ t e t, mida kuuleb või loeb.»[40]

Arutledes selle üle, mida tähendab «mõista millegi mõtet», jõuab Lindsay seisukohale, et selle omaduse üheks olulisemaks komponendiks on võime teatatust välja lugeda midagi enamat kui vahetult on antud, niisiis järelduste tuletamise võime. «Taibata rohkem kui meile teatati — see on väga iseloomulik inimlik omadus, mis avaldub enamikus seda tüüpi tegevustes, mida me nimetame arukaks.»[41]

Küsimused, mis tähendab taibata öeldu või loetu mõtet ja kuidas arvutil modelleerida järelduste tuletamise protsessi, on praegusteski keelt mõistvates süsteemides keskse kohal ning iseloomulikud nimelt neile süsteemidele. Masintõlkesüsteemides tekivad need probleemid väga piiratud ulatuses, näiteks siis, kui lause õige tõlkevariandi valimiseks on vaja mõista selle seost eelnenud või järgnevate lausetega.

Järelduste tuletusvõimet, mida R. Lindsay püüdis SAD-SAM-is modelleerida, illustreerib järgmine näide: lausest *John ja Bob mängisid aias, kui nende ema nad tuppa kutsus* peab süsteem välja lugema, et John ja Bob on vennad.

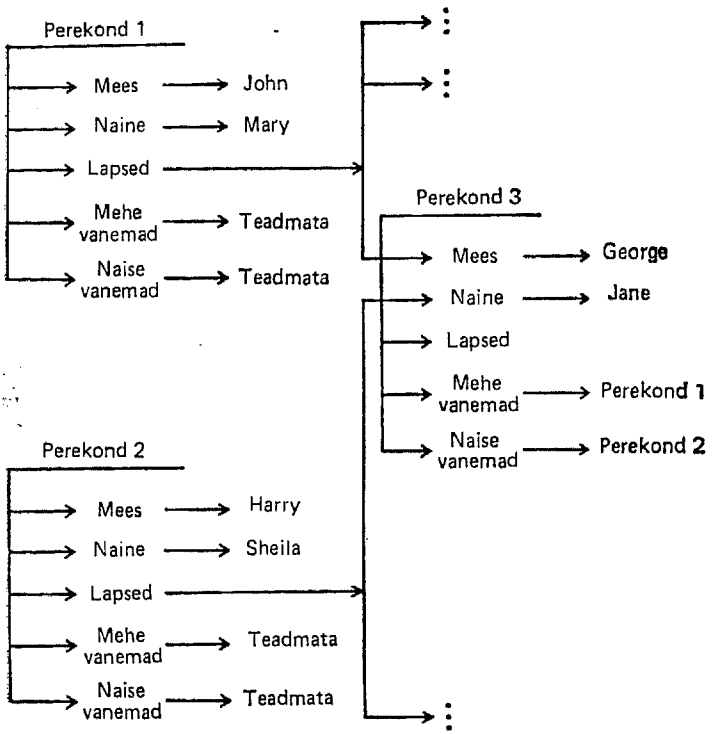
Inimeste sugulusvahekordade süsteem ongi see kitsapiiriline ala, mille piires SAD-SAM suudab järeldusi teha. Järeldused põhinevad süsteemile ette antud sugulusmõistete definitsioonidel. Sugulusvahekorrad on määratud väikese hulga lähtemõistete abil: JÄRGLANE, MEESSOOST, NAISSOOST jt. Nende abil esitatult näiteks väljend *A on B poeg* tähendab A on B OTSENE MEESSOOST JÄRGLANE. Väljendi *A on B vend* tähendust võime kirjeldada esmalt selliselt, et A ja B on samade vanemate pojad, ja edasi kasutada sõna *poeg* kirjeldust.

Nimetatud vahekordi ei valinud R. Lindsay oma süsteemi materjaliks juhuslikult. Sugulusvahekordi väljendavad mõisted on niisuguseks defineerimiseks väga sobiv materjal, mistõttu see on olnud üks populaarsemaid alasid, kus on harrastatud sõnade analüüsimist «semantilisteks elementideks».

Lause «mõistmine» SAD-SAM-is algab grammatilise analüüsiga, mille tulemuseks on lause süntaktilise struktuuri enam-vähem samasugune esitus, nagu me kirjeldasime eelmises peatükis. Sellele järgneva semantilise analüüsi ülesanne on leida lauses sugulusvahekordi tähistavad sõnad ning, kasutades nende tähenduste kirjeldusi, tuua esile ka niisugused andmed sugulusvahekordade kohta, mida lauses otseselt pole teatatud (vt. äsja toodud näidet). Andmed viiakse mällu ja säilitatakse seal eriliste struktuuridena (joonis 14),

[40] Линдсей Р. Память, способная к выводам, как основа машины, понимающей естественный язык. — В кн.: Вычислительные машины и мышление. М., 1967, с. 234.

[41] Sealsamas, lk. 237.



Joon. 14

kust on lihtne leida vastuseid küsimustele, kes on kelle isa, ema, vend jne.

Kasutades niisugust struktuuri, ei ole raske «välja arvutada» näiteks seda, kes on kelle vanaema, onu või nõbu.

Ehkki andmed, millega SAD-SAM opereeris, olid lihtsad, ilmnevad juba siin mitmed keele mõistmise modelleerimise seisukohalt olulised momendid. Esiteks peab keelt mõistev süsteem suutma vastata küsimustele, mida ei ole tekstis otseselt mainitud, kuid mis sellest järelduvad. Niisiis on tarvis reegleid järelduste tuletamiseks.

Teiseks demonstreerib SAD-SAM ilmekalt, et varjatud informatsiooni kandjaiks on ka sõnade tähendused. Seepärast tuleb keelt mõistvas süsteemis sõnade tähendused mingil viisil formaalselt kirjeldada. SAD-SAM-is kasutatakse selleks tähenduste kirjeldamist «se-

mantiliste elementide» abil. Nagu järgnevas näeme, kujuneb see valdavaks meetodiks ka hilisemates süsteemides.

Kolmandaks selgub eelnenust otseselt, et küsimustele vastamiseks kasutab süsteem kaht liiki «teadmisi»: esiteks konkreetseid andmeid Johnide, Jane'ide, Maryde jne. sugulussidemete kohta ning teiseks üldisi teadmisi selle kohta, mis tähendab olla kellegi ema, isa, mees, naine, vend või onu. Nende üldiste teadmiste hulka kuuluvad ka järeldusprotsesse juhtivad reeglid.

Kõik need kolm nõuet osutavad keele semantilise analüüsi, analüüsitavate lausete tähenduse esiletoomise ning formaalse esitamise vajalikkusele. Masintõlkesüsteemide puhul kasutatakse sõnade ja lausete semantikat puudutavaid andmeid ainult kui abimaterjali näiteks mõttetute süntaktilise struktuuri variantide välistamiseks. Keelt mõistvas süsteemis on aga, vastupidi, lause tähenduse väljaselgitamine analüüsi peaeesmärk.

Järgmiseks kirjeldame lühidalt B. Raphaeli (eespool viidatud raamatu «Mõtlev kompuuter» autori) süsteemi SIR, mis valmis 1964. aastal. Senised süsteemid nagu BASEBALL või SAD-SAM olid mõeldud opereerima vaid ülimalt kitsapiiriliste andmetega ning nendes kasutatavad andmete esitus- ja töötlemisviisid ei võimaldanud kuigivõrd olulisi üldistusi. Raphael seadis endale eesmärgiks käsitleda mõningaid andmete esitamise ja nendega opereerimise põhimõttelist laadi probleeme keelt mõistvates süsteemides.

Kõigepealt püüdis Raphael täpsemalt määratleda seda, milles peab avalduma keelt mõistva süsteemi «arusaamisvõime». Ta leidis, et keelest arusaamise parimaks demonstratsiooniks ja ühtlasi parimaks kontrolliks on dialoog, mis nõuab vestluspartnerilt arusaamist kõne all olevate andmete tähendusest ja mille käigus võib demonstreerida oma arusaamisvõimet sel teel, et annab talle esitatud küsimustele vastuseid, mida inimene hindab kui arukaid. Ei BASEBALL-i ega SAD-SAM-i puhul polnud kindlat teemat puudutav kestev dialoog ette nähtud.

Teise olulise momendina tõstis Raphael esile asjaolu, et lisaks puht keelelistele võimetele peab niisugusel süsteemil olema vestluse teemaks oleva ainevalla küllalt detailne mudel. Üks Raphaeli huvitanud põhiprobleeme seisneski selles, missugune võiks olla selles «maailmamudelis» sisalduvate faktide formaalne esitusviis, et ta ei sobiks mitte ainult mõnede konkreetsete andmetüüpide puhul, vaid oleks võimalikult üldine, seejuures aga küllalt kindla struktuuriga ning võimaldaks loogilisi arutluskäike, mida eeldab loomuliku keele lausete mõistmine.

Seetõttu ei valinud Raphael SIR-i ainevallaks mingit kindlapiirilist sisulist temaatikat, vaid abstraktsete vahekordade hulga. Vahekordadeks, millega SIR opereerib, on näiteks ülemhulga—alamhulga vahekord (*inimene — poiss*), osa—terviku vahekord (*käsi — inimene,*

sõrm — käsi), omamis- e. kuuluvusvahekord (*poisil on raamat*), ruumilise paiknemise vahekorrad (*laud on diivanist vasakul*). Raphael eeldas need olevat tüüpilised näited vahekordadest, millega inimesed arutluskäikudes opereerivad, mistõttu nende najal saab uurida küllalt laia probleemide ringi, mis puudutab teadmiste esitust inimmälus ja nende kasutamist teksti mõistmisel.

Tüüpiline dialoog inimese I ja SIR-i vahel on järgmine näide:

I: Iga poiss on inimene.
SIR: Saan aru.
I: Sõrm on käe osa.
SIR: Saan aru.
I: Inimesel on kaks kätt.
SIR: See lause on kahetähenduslik. Palun öelge teisiti.[42]
I: Kaks kätt on osa igast inimesest.
SIR: Saan aru.
I: Mitu sõrme on Johnil?
SIR: See lause on kahemõtteline. Kuid ma eeldan, et «on» tähendab «on osa». Ma ei tea, kas sõrm on Johni osa.
I: John on poiss.
SIR: Saan aru.
I: Mitu sõrme on Johnil?
SIR: Ma ei tea, mitu sõrme on käel.
I: Kummalgi käel on viis sõrme.
SIR: Saan aru.
I: Mitu sõrme on Johnil?
SIR: Kümme.

Me ei toonud seda pikka dialoogi mitte selleks, et demonstreerida SIR-i arukust või head keeleoskust. Olulisem on, et sellest dialoogist ilmneb mitu keele mõistmise protsessi iseloomustavat tõika.

Esiteks võib näha, kui loomulikult opereerib SIR loogikaliste vahekordadega. Dialoogis on kasutatud kolme vahekorda: alamhulga—ülemhulga vahekorda («iga poiss on inimene»), osa—terviku vahekorda («käel on viis sõrme») ning hulga — elemendi vahekorda («John on poiss»). SIR mitte ainult ei kasuta neid, vaid opereerib oskuslikult ka nende vastastikuste seostega. Seostades omavahel faktid, et poisid on inimesed, et inimesel on kaks kätt ja käel viis sõrme ning et John on inimene, järeldab SIR, et Johnil on kümme sõrme. Ehkki iseenesest elementaarne, demonstreerib näide seda ilmekamalt, et teksti mõistmiseks on lisaks lausete otsesest tähendusest arusaamisele tarvis ka loogikalisi arutlusi. Viimased eeldavad aga, et arutluste aluseks olevad vahekorrad, samuti nende omavahelised seosed oleksid täpselt määratletud. Niisugused loogikalis-semantilised vahekorrad lausetega väljendatud faktide vahel moodustavad teise varjatud informatsiooni liigi sõnade tähendustes sisalduva informatsiooni kõrval.

Niisama oluline on teine asjaolu,

[42] Originaalis esinev inglise keele konstruktsioon sõnaga *have* ja ka siinne eesti keele konstruktsioon sõnaga *olema* on kahetähenduslikud selles mõttes, et võivad väljendada kellegi kuulumist kui ka millegi osaks olemist. Toodud eesti keele lausel on veel kolmaski tõlgendus: «Inimese peal asetseb kaks kätt.»

mida dialoog ilmekalt demonstreerib: et tekstist arusaamiseks on süsteemil vaja piisavalt täpseid teadmisi suhtlemise objektist — maailmamudelit. Toodud näites on «maailmaks» küll ainult inimene ja ta kehaosad, kuid ka selle näite varal võib veenduda, et vastamaks meie jaoks triviaalsele küsimusele *Mitu sõrme on Johnil?* vajab süsteem teadmisi selle kohta, missugustes seostes on tegelikkuses objektid nagu sõrm, käsi, inimene, John.

SIR-i maailmamudel koosnebki objektide kirjeldustest, mille kohta talle on teatatud mingeid andmeid. Objekti kirjeldus koosneb omaduste loendist, kusjuures omaduste hulka kuuluvad ka seosed teiste objektidega. Iga omadus kujutab endast põhimõtteliselt samasugust paari «tunnus—väärtus» nagu BASEBALL-is. SIR-is on see esitusviis kasutusel aga märksa üldistatumal kujul, samuti on välja töötatud üldised protseduurid niisuguste kirjeldustega opereerimiseks, näiteks uue informatsiooni lisamiseks mällu või vajalike andmete mälust otsimiseks. Lause analüüsi eesmärgiks süsteemis on teisendada see kujule, mis määrab vastavad kirjelduste töötlemise protseduurid. Lause tähendust võibki määratleda kui töötlemisprotseduuri, mille see lause kindlaks määrab. Näiteks lause *Poisid on inimesed* tähendus SIR-i jaoks avaldub protseduurides, mis lisavad objekti (mõiste) «poiss» kirjeldusse mälus omaduse ÜLEMHULK = *inimene,* mõiste «inimene» kirjeldusse aga omaduse ALAMHULK = *poiss.* Lause mõistmise kui protsessi formaalseks mudeliks on lausele vastavate protseduuride teostamine. Mudeli adekvaatsust võime hinnata selle järgi, missugusel määral keelt mõistva süsteemi mälus kujunenud objektide kirjeldused vastavad neile teadmistele vastavate objektide kohta, mida analüüsitavast lausest saab inimene.

1960-ndatest aastatest pärineb rohkesti teisigi keelt mõistvaid süsteeme. Loodi ja katsetati erinevaid süsteemitüüpe, osa neist tuginesid samasugustele või lähedastele põhimõtetele nagu äsja kirjeldatud SIR. Näiteks D. Bobrow' poolt 1964. aastal loodud süsteem STUDENT lahendas sõnalisi algebraülesandeid, mis olid esitatud loomulikus keeles, nagu näiteks järgmine: «Mary oli kaks korda vanem Tomist, kui Mary oli nii vana nagu Tom praegu. Mary on praegu 24-aastane. Kui vana on Tom?» Teksti põhjal koostas STUDENT vajaliku võrrandi ja lahendas selle.

Loodi aga ka hoopis teistsugustest põhimõtetest lähtuvaid süsteeme. Ühe niisuguse süsteemina võib nimetada ELIZA-t, millest oli juttu peatüki alguses. ELIZA autorit J. Weizenbaumi ei huvitanud niivõrd üksikute lausete mõistmine, kuivõrd inimestevahelist vestlust tervikuna juhtivad seaduspärasused. Nende seaduspärasuste modelleerimiseks valis ta omapärase vormi: ELIZA jäljendas psühhiaatrit, kes vestleb patsiendiga; viimase rollis esines süsteemiga vestlev inimene.

Weizenbaumi põhimõtteid arendas hiljem edasi J. Colby, konst-

rueerides süsteemi, kus inimese ja süsteemi rollid olid vahetatud. Tema süsteem kujutas endast paranoiku mudelit, mille suhtlemisviis pidi jäljendama paranoiku mõtlemis- ja suhtlemisstereotüüpe. «Tehisparanoiku» vestlusest inimesega pärineb ka käesoleva peatüki motos toodud dialoog. Paraku pakub Colby süsteem huvi pigem teatud psühholoogiliste mehhanismide mudelina kui keele mõistmise protsessi mudelina.

Võrdlemisi tuntuks sai ka M. Ross Quilliani poolt katsetatud «semantilise mälu mudel». Ka selles süsteemis oli raskuspunkt teadmiste korraldamise psühholoogiliste mehhanismide modelleerimisel. Kuid Quilliani huvitasid need mehhanismid nimelt keele mõistmise pühholoogilise alusena. Tema semantiline mälu kujutab endast mitmesuguste seoste kaudu ühendatud semantiliste, s. o. tähenduslike üksuste võrku. Erilist tähelepanu pööras Quillian semantiliste üksuste suhtelisusele: struktuur, mida ühes kontekstis vaadeldakse kompleksena, võib teises kontekstis talitleda kui iseseisev üksus, mida tervikuna iseloomustavad kindlad omadused. Inimestele on teadmiste niisugune organisatsioon väga iseloomulik. Näiteks avaldub see võimes oma teadmiste süsteemis defineerida ja kasutusele võtta kui tahes keerulisi mõisteid ning seejärel opereerida nendega kui terviklike, jagamatute elementidega. Quilliani ideed osutusid koheseks realiseerimiseks ilmselt liiga keerulisteks, igatahes ei loodud ühtki neil põhinevat keelt mõistvat süsteemi. Quillian ise katsetas semantilise mälu mudeliga, mis ei analüüsinud ega sünteesinud lauseid, vaid ainult võrdles talle esitatud mõisteid ning püüdis vastavatest definitsioonidest lähtudes kehtestada nende vahel kindlaid seoseid.[43]

Kirjeldatud süsteemide najal sai lugeja usutavasti ettekujutuse sellest, kuidas algas süsteemide loomine, mis peavad modelleerima inimese intellektuaalsete võimete üht avaldust — võimet mõista loomulikku keelt. Siinjuures ilmnesid eos kõik olulisemad probleemid, mis on niisugustele keeletöötlussüsteemidele eripärased. Järgnevalt asume lähemalt vaatlema, milles need probleemid seisnevad ja missuguste vahendite abil neist püütakse üle saada.

4.3. Semantiline analüüs ja teadmiste semantilised mudelid

«On tulnud aeg, kus me lõpuks asume tegelema probleemiga, millega võrreldes kõik selles raamatus seni kirjeldatud seadmed tunduvad lausa mänguasjadena. Me teesklesime, nagu ei märkaks üldse seda sügavikku, kuhu nüüd oleme

[43] M. Ross Quilliani semantilise mälu ehitust on lühidalt kirjeldatud raamatus H. Õim. Semantika. Tln., 1974, lk. 158 jj. Sealsamas on põgusalt tutvustatud ka mõningaid teisi kõnealusest perioodist pärinevaid süsteeme.

otsustanud pilgu heita. Räägime kõike seda selleks, et valmistada lugejat ette kohtumiseks mõistatusega, mis on salapärasem kui sfinks, ja nimelt tähenduse probleemiga, kõige põhjatumaga kõigi põhjatuste seas. Igaühe kohal, kes tegeleb keelega kui spetsialist, ripub Damoklese mõõgana tähenduse probleem.»[44]

Niisuguse mõjuva sissejuhatusega alustab Stanislaw Lem oma «intellektroonikaalastes» arutlustes küsimust, kuidas inimesed mõistavad nende keeleväljendite mõtet, mida nad kasutavad omavahelises suhtlemises, ja kuidas seda mõistmist reprodutseerida intellektuaalsetes automaatides.

Möödunud poolteise aastakümne jooksul on tähenduse olemuse salapära küll märksa vähenenud, kuid väited tähenduse probleemi keerukusest ja olulisusest, nii nagu ka väide tähenduse «põhjatusest» on praegu samavõrd paikapidavad ja aktuaalsed, nagu nad olid seda S. Lemi raamatu ilmumise ajal. Oleme õppinud formaliseerima ja elektronarvutil modelleerima paljusid semantilisi, s. o. tähendusse puutuvaid protsesse. Kuid see on vaid kübe, võrreldes selle tohutu osaga, mida märgid, need tähendust omavad üksused ja struktuurid, etendavad inimeste elus.

Dramatism, millega Lem toob oma arutlustesse tähenduse mõiste, kajastab viisi, kuidas see mõiste tuli nii keele teoreetilistesse käsitlustesse kui ka keele automaattöötlusse. Pikka aega nähti tähenduses nähtust, mida iga täpsust armastav teaduslik distsipliin ja ka rakendusala peab vältima. Keeleteaduses valitses see seisukoht 1950-ndatel aastatel näiteks strukturaallingvistikas, mis oli keele automaattöötluse teoreetiliste ideede põhiline allikas. Tähenduse vältimiseks mõeldi välja ülikeerulisi ja teinekord ka ülimalt kavalaid analüüsivõtteid. Kuid need formaalsed protseduurid ei saanud uurijail kuigivõrd aidata mõista, kuidas «loomulik keel saadab korda oma igapäevaseid imesid».

Tõdemus, et ei ole vaja tähendusest lahti saada, vaid on vaja õppida temaga ümber käima, põhjustas faktiliselt pöörde keele (ja mitte ainult keele) automaattöötluses. Järsk huvi tõus keelt mõistvate süsteemide loomise vastu 1960-ndate aastate teisel poolel oli selle pöörde üks avaldusi.

Küsimuse kohta, mis on tähendus, on kirjutatud terveid raamatukogusid. Selle kallal on juurelnud filosoofid, loogikud ja muidugi keeleteadlased. Intellektuaalsete infosüsteemide loojatele ei ole oluline niivõrd välja selgitada, mis on tähendus, kuivõrd leida viis, kuidas reprodutseerida elektronarvutis protsesse, milles avaldub tähendus. Tänu ülesande niisugusele sõnastusele ollakse olukorras, kus tähenduse olemuse selgitamine kulgeb rööbiti keelt mõistvate süsteemide loomisega, mitte aga pole selle eelduseks. Seetõttu ei hakka ka meie praegu semantika teoreetilistel probleemidel pikemalt peatuma, vaid asume vaatlema keelt mõist-

[44] Лем С. Сумма технологии. М., 1968, с. 206.

vate süsteemide «semantilist varustust».

Keelt mõistva süsteemi esimene põhiline struktuurne erinevus näiteks masintõlkesüsteemist avaldub selles, et tema koosseisu kuulub lisaks morfoloogia- ja süntaksikomponendile tingimata ka semantikakomponent, mille ülesanne on teisendada lause süntaktilise analüüsi tulemusena saadud süntaktiline esitus (sellest rääkisime eelmises peatükis) semantiliseks esituseks. Lause semantiline esitus — see on tema tähenduse, mõtte formaalne kirjeldus. Mainitud teisendusprotseduuri nimetatakse lause semantiliseks analüüsiks. Selleks peab semantikakomponent sisaldama vähemalt kaht liiki formaalseid vahendeid: esiteks semantilise analüüsi reegleid ja teiseks vahendeid analüüsi tulemusena saadud semantiliste struktuuride kirjapanemiseks ja nendega opereerimiseks.

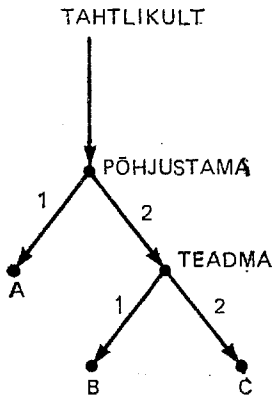
Lause semantiline esitus moodustub tähenduselemente esitavatest semantilistest üksustest ja nendevahelistest seostest. Lause süntaktilise struktuuri üksused on sõnad. Semantiliste üksuste olemus on märksa ebaselgem. Enamik sõnu on tähenduslikult komplekssed. Selleks et esile tuua nende tähendustes sisalduvat varjatud informatsiooni, on vaja ka sõnade tähendused esitada kindlate struktuuridena, põhimõtteliselt niisamuti, nagu seda tehakse näiteks seletavates sõnaraamatutes. Ainult et tähenduste kirjeldused peavad olema vormistatud täpselt määratletud formaalsete reeglite järgi.

Üksusi, millest tähenduste kirjeldused koostatakse, on nimetatud mitmeti — semantilisteks elementideks, elementaartähendusteks, semantilisteks primitiivideks (algüksusteks). Meie nimetame neid edaspidi elementaartähendusteks, mõistes nende all üksusi, mille tähendusi pole enam vaja jagada väiksemateks üksusteks.

Et sõnade tähendused pole jagamatud tervikud ja et peavad eksisteerima lihtsamad, elementaarsemad tähenduslikud üksused, seda tõendavad nii keeleteaduse andmed kui ka vastavad psühholoogilised uurimused. Kuid elementaartähenduste olemuse täpsema määratlemise ümber kestavad ägedad diskussioonid. See on aga juba rohkem teoreetilise keeleteaduse ja psühholoogia probleem, millel me siin ei pruugi peatuda. Põhimõtteliselt võime käsitleda elementaartähendusi kui lihtsaid, semantiliselt jagamatuid sõnu.

Elementaartähenduste kasutamist sõnade ja edasi ka lausete tähenduste kirjeldamisel on kõige lihtsam selgitada mõne näite varal. Võtame näiteks sellise lihtsa sõna kui *teatama*.

Kui me teame, et keegi (A) teatas kellelegi (B-le) midagi (C), siis tegelikult teame mitte üht, vaid mitut fakti, kusjuures nende teadmine ilmneb näiteks selles, et me järeldame need faktid väitest, et A teatas B-le, et C. Nii järeldame, et B teab (nüüd), et C, ja järeldame, et see teadmine tekkis tal A tegevuse tulemusena — A põhjustas selle teadmise; edasi järel-



Joon. 15

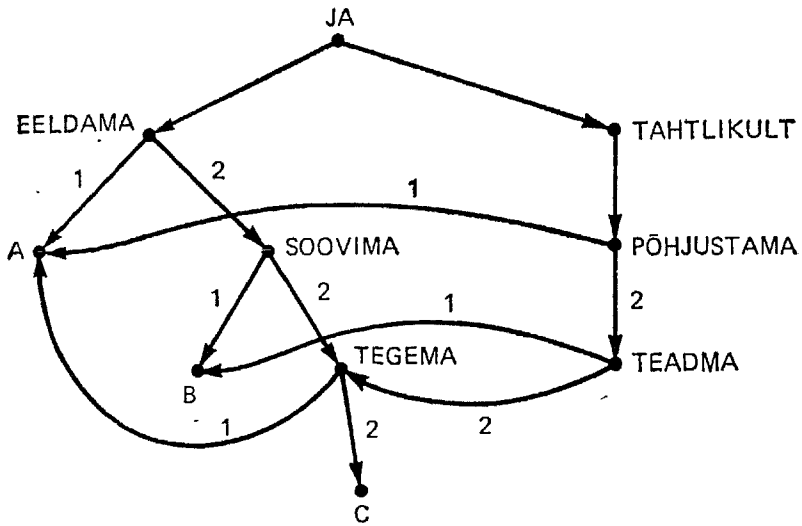
dame samuti, et A tegevus oli tahtlik (kui A põhjustanuks vastava teadmise B-l kogemata, siis ei nimetaks me seda tegevust teatamiseks, vaid näiteks väljalobisemiseks). Niisiis oleme sõna *teatama* tähenduse kirjeldamisel kasutanud üksusi PÕHJUSTAMA, TEADMA, TAHTLIKULT. Nende üksuste omavahelisi seoseid võime kirjeldada struktuuriga joonisel 15.

See struktuur on võrdlemisi sarnane eelmises peatükis käsitletud süntaktiliste struktuuridega. Pisut keerulisema struktuuri näitena võib tuua sõna *tõotama (A tõotab B-le, et teeb C)* tähenduse esituse joonisel 16.

Selle struktuuriga esitatavat sisu võib kirjeldada järgmiselt: A eeldab, et B soovib, et ta (A) teeks C, ja (seda eeldades) teatab B-le, et teeb C. Üksusest JA lähtuv parempoolne haru viib *teatama* struktuuri juurde, vasakpoolne haru viitab A eelduse sisule. Väliselt see struktuur enam ei sarnane lausete süntaktilisi struktuure esitavate graafidega. Seda ühel põhjusel: ühte ja samasse üksusesse suundub siin mitu noolt, süntaktilistes struktuurides on selline olukord üldreeglina välistatud.

Teatud määral taandub see erinevus muidugi tehniliseks väljajoonistamise erinevuseks. Kuid me kasutasime sellist esitusviisi meelega, juhtimaks tähelepanu ühele põhimõttelisele erinevusele süntaktiliste ja semantiliste esituste vahel. Süntaktilise esituse alus on lause oma lineaarselt järjestatud üksustega — sõnadega. Lineaarsus ei ole ainult lause välist pilti puudutav iseärasus, vaid temaga on seotud mitmed lause süntaktilist organisatsiooni iseloomustavad seaduspärasused, nagu näiteks süntaktilise struktuuri projektiivsus, samuti asjaolu, et üks sõna ei saa üldreeglina esineda kahe või rohkema sõna laiendina; just seetõttu suundub süntaktilises esituses igasse sõnasse vaid üks nool. Sõnade kui ka fraaside semantilised struktuurid ei ole enam seotud niisuguste kitsendustega. Semantiline struktuur, nagu meie teadmistegi struktuur, on «ruumiline», see esitab paljusid samaaegselt kehtivaid väiteid. Kuivõrd teatav üksus (nagu näiteks A eeltoodud struktuuris) esitab üht kindlat reaalsuse objekti, nii esineb ta selles struktuuris vaid üks kord, ning asjaolu, et tema kohta kehtib korraga mitu erinevat väidet, avaldubki selles, et temasse suunduvad nooled mitme üksuse juurest.

Seniöeldu puudutas sõnade tä-



Joon. 16

henduste esitust. Kuid nüüd võime öelda, et lausete semantilised esitused ei erine esitusviisi poolest põhimõtteliselt sõnade semantilistest esitustest. Ka lausete semantilised esitused moodustuvad elementaartähendustest ja nendevahelistest seostest. Näiteks lause *Pikkade juustega kõhn noormees tõotas mossitavale tütarlapsele, et ta toob kuu taevast alla* semantilise esituse erinevus eeltoodud struktuurist on selles, et A asemel on siin semantiline struktuur, mis vastab tervele väljendile *pikkade juustega kõhn noormees,* B asemel väljendi *mossitav tütarlaps* struktuur, C asemel aga struktuur, mis vastab lausele *Ta (=poiss) toob kuu taevast alla.* Tuleb rõhutada, et sõnade ja lausete semantilised esitused on sarnased nimelt formaalse esitusviisi poolest. Lausete semantilise struktuuri kujutamisel on mitmeid probleeme, mis sõnade puhul üldse ei teki. Sõna struktuur on «kord ja alatiseks» fikseeritud (muidugi teatud suhtelistes piirides), see on tükike meie teadmisi, millega me opereerime kui tervikuga, olgugi tal teinekord väga keeruline sisestruktuur. Lausete puhul tekivad probleemid just seetõttu, et sõnade erinevate seostumiste korral võime saada hoopis erinevad struktuurid.

Seda tüüpi struktuure, mida semantiliste esitustena illustreerivad eespool toodud graafid (eriti viimasena esitatu) on hakatud nime-

tama semantilisteks võrkudeks. Graafi tippudeks on semantilised üksused (elementaartähendused), kaarteks aga suunatud ja kindlal viisil tähistatud seosed üksuste vahel (neist seostest pole me seni rääkinud, kuid neil peatume allpool pikemalt).

Semantilised võrgud on käesoleval ajal kõige levinum semantilise esituse formaliseerimisviis. Selle kõrval on teine kasutatavam esitusviis loogikast ülevõetud predikaatarvutuse formalism, kus sõnade ja lausete tähendused esitatakse predikaatidest ja argumentidest moodustatud struktuuridena. Näiteks väljendi *A tõotas B-le, et teeb C* tähenduse esitus näeks predikaatarvutuse formalismis välja järgmiselt:

JA (EELDAMA (A, SOOVIMA (B, TEGEMA (A, C))), TAHTLIKULT (PÕHJUSTAMA (A, TEADMA (B, TEGEMA (A, C)))))).

Predikaatarvutuse vormelite kasutamine semantiliste esitustena levis keelt mõistvates süsteemides kahel põhjusel. Esiteks kasutati 1960-ndate ja 70-ndate aastate vahetusel seda esitusviisi laialt lingvistilises semantikas. Teiseks — ja see oli ilmselt peamine — kasutati sedasama vormi andmete esitamiseks ka teistes tehisintellektisüsteemide tüüpides ning selle formalismi jaoks olid välja töötatud mitmesugused andmetega opereerimise protseduurid. Semantiliste võrkude esitusviis on siiski osutunud paindlikumaks ja adekvaatsemaks vahendiks inimteadmiste mitmekesiste struktuuride edasiandmisel, mistõttu see pole tänapäeval levinud mitte ainult keelt mõistvates süsteemides, vaid ka teistes tehisintellektisüsteemides. Allpool semantilistest esitustest rääkides peame silmas eelkõige semantilisi võrke.

Elementaartähendused on üksused sõnade ja lausete tähendusi esitavates semantilistes võrkudes. See ei ole semantiliste võrkude kohustuslik tunnus. Nagu allpool näeme, ei või neiks üksusteks olla mitte ainult teadmiste elementaartasandi, vaid mingi kõrgema komplekstasandi elemendid, kui vaid nende omadused ja omavahelised seosed on küllaldaselt määratletud. Kirjeldades näiteks mõne jutustuse kui terviku sisulist struktuuri, ei ole sugugi otstarbekas püüda seda kohe esitada elementaartähenduste kaudu. Sel tasemel osutuks struktuur nii keeruliseks, et meil kaoks võimalus näha või väljendada mingeid üldisemaid seaduspärasusi. Kui aga valime esimese esitustasandi üksusteks «tegelaskujud», «tegelaste eesmärgid», «sündmused» jms., siis avaldub jutustuse struktuur märksa selgemini.

Elementaartähenduste tasand on see, kuhu lauseid analüüsides on vaja välja jõuda. Keelt mõistvad süsteemid on enamasti loodud mingit kitsamat temaatikat käsitlevate tekstidega opereerimiseks. Seepärast pole põhjust oodata neilt elementaartähenduste kuigivõrd universaalseid loendeid. Mõnel näitel tasub siiski peatuda.

Keelt mõistvate süsteemide loojate tähelepanu on paelunud tegevusi väljendavad sõnad (rõhutame — mitte tegusõnad, vaid tegevusi tähistavad sõnad; *jooksmine, hüpe* või *naer* tähistavad tegevusi, ehkki on nimisõnad, kuna *õitsema* või *haisema* ei tähista tegevusi, ehkki on tegusõnad). Nende sõnade tähenduste kirjeldamiseks on pakutud mitmeid elementaartähenduste süsteeme. Üht kõige radikaalsemat lähenemisviisi esindab R. C. Schank, keelt mõistvate süsteemide üks tuntumaid teoreetikuid USA-s (tema üldistest seisukohtadest keele mõistmise modelleerimisel tuleb allpool lähemalt juttu). Schank leiab, et kõnealuste sõnade tähendused on võimalik esitada väga väikese arvu elementaartähenduste — «elementaartegude» abil. Nende arv on erinevates käsitlustes kõikunud, kuid kunagi pole see olnud suurem viieteistkümnest. Schanki elementaartegude tüüploend on järgmine.

1. PTRANS («physical transfer») — materiaalse eseme liigutamine, viimine ühest kohast teise. Esineb sõnades, nagu *minema, lendama, tõstma, viskama.*

2. ATRANS («abstract transfer») — abstraktse omaduse või suhte, näiteks omamissuhte ülekandmine: *andma, võtma, müüma* jt.

3. MTRANS («mental transfer») — informatsiooni ülekandmine ühelt olendilt teisele (või ka olendi «sees»): *teatama, käskima, meenuma, unustama* jt.

4. MBUILD («mental building») — uute infostruktuuride, nagu ideede, plaanide jms. konstrueerimine mälus leiduvast informatsioonist: *mõtlema, taipama, otsustama.*

5. PROPEL — füüsilise jõu rakedamine mingile objektile: *lööma, tõukama.*

6. INGEST — mingi objekti «viimine» elusolendi organismi teatud kindlal viisil: *sööma, jooma, hingama* jne.

7. EXPEL — objekti (või ka heli) «väljaviimine» organismist: *köhima, aevastama, oksendama.*

8. MOVE — (passiivne) liikumine: *voolama, kukkuma, hajuma.*

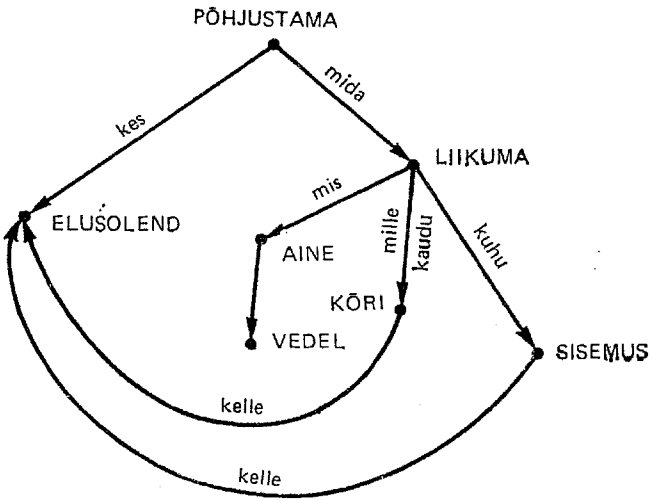
9. GRASP — mingi objekti haaramine: *kahmama, (kätte) võtma.*

10. ATTEND — meeleorgani või tähelepanu suunamine millelegi: *vaatama, nuusutama, maitsma, jälgima.*

11. SPEAK — kõnehäälitsuste tegemine: *kõnelema, hüüdma, kutsuma.*

Neist elementidest konstrueeritakse kõigi tegevusi väljendavate sõnade tähendused. Niisuguses «maksimalistlikus» lähenemises pole Schankil olnud eriti palju järgijaid. Schanki kaitseks võib öelda, et ta **on kirjeldatud** elementaartegude süsteemi (koos elementaartähendustega seisundite, omaduste, vahekordade jms. kirjeldamiseks) kasutanud rohkem kui ühes keelt mõistvas süsteemis ning seda küllalt edukalt.

Kõrvutamiseks võib tuua Y. Wilksi, keelt mõistvate süsteemide teise tuntud teoreetiku poolt tegevuste kirjeldamiseks kasutatud ele-



Joon. 17

mentaartähenduste loendi. Wilks kasutab oma elementaartähenduste esitamiseks sobivaid inglise keele sõnu. Selguse huvides esitame siin nende eestikeelsed tõlked, ehkki sageli ligikaudsed: VALIMA, PÕHJUSTAMA, MUUTUMA, TUNDMA, OMAMA, MEELDIMA, MAKSMA, KASUTAMA, SOOVIMA, RÄÄKIMA, OLEMA, TEGEMA, LIIKUMA, ÜMBRITSEMA, VOOLAMA, KUKKUMA, TOIMUMA. Näiteks sõna *jooma* tähendust võib neid elementaartähendusi kasutades kirjeldada nii nagu joonisel 17 (Wilksi enda näide, ehkki tema poolt kasutatava esitusviisi «graafiline pilt» on pisut teine). Seda struktuuri lugedes saame umbes väljendi *elusolend põhjustab vedela aine liikumise oma kõri kaudu oma sisemusse.*

Ka nende kahe näite põhjal peaks olema selge, et elementaartähenduste loendite esitamisel pole suurt informatiivset väärtust. Need loendid on eri autoritel erinevad, mis muu hulgas tõendab, et empiiriliselt ollakse elementaartähenduste mingi kindla ja lõpliku, keemiliste elementide tabeli taolise süsteemi väljaselgitamisest kaugel.

Elementaartähendustest põhimõtteliselt rääkides on aga vaja esile tuua järgmist. Esiteks, ei maksa eriti tähelepanu pöörata asjaolule, et elementaartähenduste kaudu esitatud sõnade tähenduste kirjeldused mõjuvad ebaharilike ja isegi ebaloomulikena. Elementaartähenduste kasutamise eesmärk ei ole sõnade tähenduste varustamine inimlikult loomulike definitsioonidega, vaid selles, et esile tuua var-

jatud informatsioon, mis sisaldub sõna tähenduses ja mida inimene sellest tegelikult välja loeb. Võib tunduda kummalisena kirjeldada *jooma* tähendust nii, nagu seda teeb Wilks. Kuid fakt on see, et kui me mingis lauses loeme midagi joomisest (esineb sõna *jooma* vastavas vormis), siis me teame ka, et see, kes jõi, oli elusolend (*jooma* algse tähenduse puhul), et see, mida joodi, oli vedelik, et see vedelik liikus ja seejuures liikus nimelt jooja sisse tema kõri kaudu.

Võib muidugi küsida, kas kogu see teadmine «joomise anatoomiast» on nii väga oluline ühele keelt mõistvale süsteemile. Kuid see on juba teist laadi küsimus. Mis on keelt mõistvale süsteemile oluline, see sõltub süsteemi otstarbest. Tegelikult ei analüüsita üheski süsteemis elementaarstruktuurideks kõiki sõnu, millega süsteem opereerib, vaid ainult neid, milledes võib sisalduda süsteemi eesmärkide seisukohalt olulist informatsiooni. Näide sõnaga *jooma* on hea just sellepärast, et võimaldab olulisuskriteeriumi selgelt esile tuua.

Elementaartähenduste teine põhimõtteline omadus puudutab nende kui semantiliste üksuste rolli teadmisstruktuuride formeerimisel. Elementaartähendused ei kombineeru omavahel juhuslikult. Iga elementaartähendust iseloomustab kindel arv harusid, mille kaudu ta saab teistega seostuda. Need võimalikud harud on elementaartähenduste samasugune iseärasus nagu süntaktilisel tasandil sõnade süntaktiline valents, iseloomustades antud üksuse võimet moodustada ühendeid teiste üksustega. Igal harul on kindel suund ja sisu, mis määratleb teise üksuse rolli antud lähteüksuse poolt tähistatavas situatsioonis või tegevuses. Niisiis mitte ainult elementaartähendused üksi, vaid ka nendevahelised seosed osalevad struktuuri kui terviku poolt väljendatava sisu loomises.

Kirjeldatud omadus ei ole iseloomulik mitte ainult elementaartähendustele, vaid suvalise tasandi semantilistele üksustele — see on semantiliste võrkude kui esitusviisi üks iseloomustavaid jooni.

Semantiliste seoste süsteeme on nii keeleteaduslikus kui ka keelt mõistvate süsteemide alases kirjanduses välja pakutud ja praktikas kasutatud mitmeid, niisamuti kui elementaartähendustegi süsteeme. Keelt mõistvate süsteemide puhul on oluline faktor jällegi süsteemi otstarve ja tekstide ainevaldkond. Kõigil pakutud loenditel võib täheldada aga teatud ühisosa, seoste hulka, mille olulisust tunnistab enamus uurijaid. Esitame semantiliste seoste ühe «näidisloendi», mis haarab kõige tüüpilisemaid seoseliike.

1. AGENT — seos (haru), mis lähtub tegevust tähistavast üksusest ja seob seda tegevuse sooritajat tähistava üksusega; tähistame lühidalt tähisega AG. Näiteks:

AG
poiss karjub : karjuma → poiss,
kingiti kaastööliste poolt : kinkima AG *→ kaastöölised.*

2. OBJEKT — viitab tegevuse

poolt mõjutatavale objektile (tähis OBJ):

OBJ

kinkis raamatu : *kinkima* → *raamat.*

3. RETSIPIENT — viitab olendile, kes tegevuse tagajärjel midagi «saab» (tähis RET):

RET

kinkis tütrele : *kinkima* → *tütar,*

RET

teatas sõbrale : *teatama* → *sõber,*

RET

käskis poissi : *käskima* → *poiss.*

4. PATSIENT — seisundi või omaduse passiivne «kandja», «taluja» (tuleneb ladinakeelsest väljendist *paciens* 'kannataja'; tähis PAT):

PAT

poiss on õnnetu : *õnnetu* → *poiss,*

PAT

poiss väsis : *väsima* → *poiss.*

5. INSTRUMENT — viitab tegevuse sooritamise vahendile (tähis INST):

INST

lõi (kedagi) kepiga : *lööma* → *kepp.*

6. LÄHTEKOHT — viitab kohale, kust liikumine algab (tähis LK):

LK

tuli toast : *tulema* → *tuba.*

7. SIHTKOHT — viitab kohale, kuhu liikumine suundub (tähis SK):

SK

viis õue : *viima* → *õu.*

8. KOHT — viitab tegevuse toimumise või situatsiooni kehtimise kohale (tähis K):

K

jooksis tänaval : *jooksma* → *tänav.*

9. EESMÄRK — olukord või situatsioon, mis on antud tegevuse eesmärgiks, otstarbeks (tähis EM):

jooksis (midagi) vaatama
jooksis, et vaadata } :

EM

jooksma → *vaatama*

10. RESULTAAT — olukord või situatsioon, mis on antud tegevuse tulemuseks või tagajärjeks (tähis RES):

RES

muutus rõõmsaks : *muutuma* → *rõõmus.*

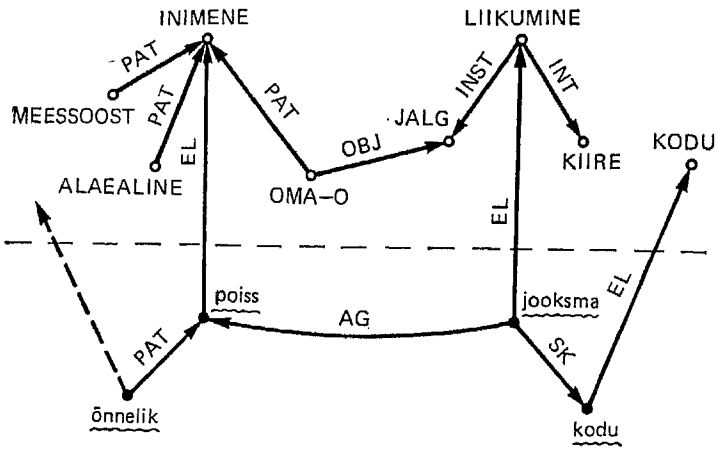
11. INTENSIIVSUS — tegevuse protsessi või seisundi intensiivsus (tähis INT):

INT

rääkis valjusti : *rääkima* → *valjusti.*

Iga elementaartähendust iseloomustab kindel kombinatsioon temast lähtuvaid seoseid (formaalselt: temast väljuvaid harusid, mis on tähistatud vastavate seoste tähistega), mis moodustavad tema semantilise valentsi. Sõna tähendusse kuuluvate elementaartähenduste kaudu omandab sõna teatud semantilise valentsi; lihtsalt öeldes moodustavad selle need harud, mis elementaartähenduste antud sõnaks ühendamisel on vabaks jäänud.

Kirjeldatud vahekordadele semantiliste üksuste ja struktuuride vahel tuleb aga lisada veel teist liiki sidemed. Senikirjeldatud on kõik süntagmaatilised seosed, mis seovad üksusi nende semantiliste laienditega. Ent teadmistes on üksused seotud veel teist liiki, nimelt hierarhiliste ehk paradigmaatiliste seostega. Neist olulisimad on ülemhulga-alamhulga seos ehk sooliigiseos, millest oli palju juttu 2. peatükis, ning elemendi-hulga- ehk hulkakuuluvusseos. Tähistame



Joon. 18

esimest seost SUP-iga (viitab ülem- ehk supermõistele), teist SUP EL-ga: INIMENE → ELUSOLEND (inimesed on üks liik elusolendeid); JÜRI → INIMENE (Jüri — keegi konkreetne inimene nimega Jüri — on kategooria INIMENE üks esindaja). Esimest nimetatud kahest paradigmaatilisest seosest läheb vaja, et osutada vastavaid vahekordi semantiliste üksuste või struktuuride vahel mälus. Teist seost on vaja lausete semantilistes esitustes seetõttu, et enamasti ei ole tekstides juttu mitte üldmõistetest (inimesest üldse, jooksmisest üldse), vaid nende mõistete üksikesindajatest, s. t. konkreetsetest inimestest, konkreetsest jooksmisaktist.

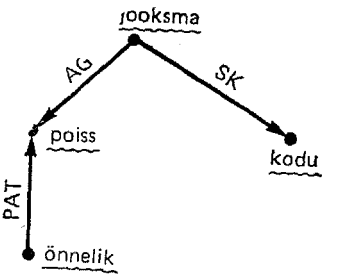
Kujutame nüüd illustratsiooniks lause *Õnnelik poiss jooksis koju* ligilähedast semantilist esitust joonisel 18.

Selles on sõna *poiss* tähendus esitatud kui «alaealine meessoost inimene» ja *jooksma* tähendus kui «kiire liikumine oma jalgade abil». Neid tähendusi esitavad struktuurid moodustuvad elementaartähendustest, mis esindavad üldisi kategooriaid (üksus OMA-O väljendab mõistet «omab kui osa»). Lauses räägitakse aga ühest konkreetsest poisist ja tema jooksmisest oma koju. Niisugustele üksikobjektidele ja -tegevustele peavad vastama keelt mõistva süsteemi teadmistes omad üksused, mille juurde koonduvad nende individuaalsed omadused. Seetõttu moodustub neis teadmistes otsekui kaks kihti, kaks teadmisstruktuuride piirkonda — ühte kuuluvad teadmised üldmõis-

tetest ja üldistest seaduspärasustest, teise teadmised konkreetsetest individuaalsetest objektidest, sündmustest ja olukordadest kui esimeste «üksikeksemplaridest». Jaotus on teatud määral küll tinglik, eriti näiteks omadusi, seisundeid või tegevusi väljendavate mõistete puhul. Kuid põhimõtteline vajadus niisuguse jaotuse järele on siiski selge.

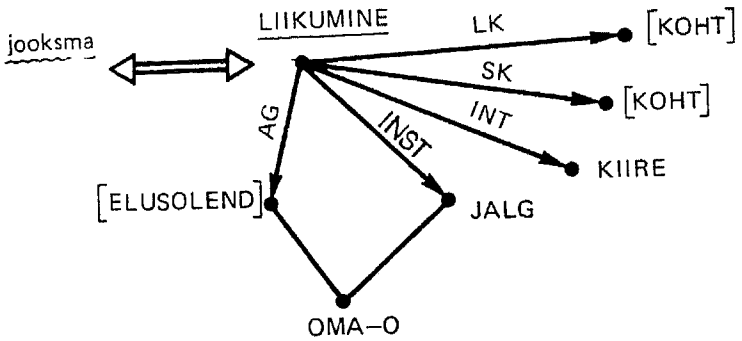
Nüüd võime pöörduda otseselt semantilise analüüsi juurde, mille ülesanne on teisendada lause süntaktiline esitus semantiliseks esituseks.

Erinevalt morfoloogilisest ja süntaktilisest analüüsist ei ole semantilises analüüsis selgeid, kindlakskujunenud meetodeid. Selle põhjuseks on erinevused üksikute autorite poolt kasutatavates semantilise esituse vormides, aga ka asjaolu, et keelt mõistvates süsteemides mõistetakse sageli erinevalt ka süntaksi ja semantika vahekorda. Kirjeldame alljärgnevalt, kuidas põhimõtteliselt kulgeb semantiline analüüs eeldusel, et lausete süntaktilised esitused kujutavad endast selliseid sõltuvusstruktuure, nagu me kirjeldasime eelmises peatükis. Sel juhul seisneb semantilise analüüsi esmane ülesanne selles, et lause süntaktiline sõltuvusstruktuur ümber kujundada vastavaks semantiliseks võrguks. Vaatleme protsessi sellesama lause *Õnnelik poiss jooksis koju* põhjal, mille semantilise esituse me tõime. Lause süntaktiline esitus näeb välja nii nagu joonisel 19 (morfoloogilise informatsiooni jätame kõrvale).



Joon. 19

Peamise lähtematerjali vajalikeks teisendusteks saab sõnastikust. Keelt mõistva süsteemi sõnastikus on sõna tähendust kirjeldav «semantiline tsoon» märksa ulatuslikum kui masintõlkesüsteemi sõnastikus, kus ta piirdus sõna üldiste semantiliste tunnuste (semantiliste kategooriate) loendiga. Keelt mõistva süsteemi sõnastikus on sõna semantilises tsoonis toodud tema tähenduse terviklik kirjeldus, tema semantiline esitus, mille näiteid tõime eespool. Siin võib lisada, et analüüsi seisukohalt vaadatuna võib sõna semantilist esitust tõlgitseda kui reeglit, mis ütleb, et antud sõna tuleb lause semantilisele esitusele üle minnes asendada nii- ja niisuguse struktuuriga. Sünteesi puhul on asendusprotsess aga vastupidine, nii et reegel on kahesuunaline. Näiteks sõnadel *jooksma* ja *poiss* peaksid sõnastikus leiduma sellised semantilised esitused-reeglid, nagu näha joonistel 20 ja 21. Nende esituste kohta on vaja veel teha paar selgitavat märkust.



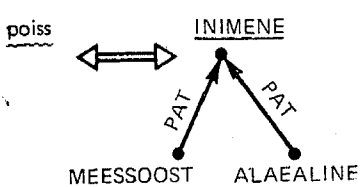
Joon. 20

Iga sõna tähenduses on domineerivaks üksuseks üks kindel kategooria, teised täpsustavad tema sisu vastavatest aspektidest. Sõnal *jooksma* on selleks LIIKUMINE, sõnal *poiss* ELUSOLEND. Toodud struktuurides on neile kategooriatele joon alla tõmmatud.

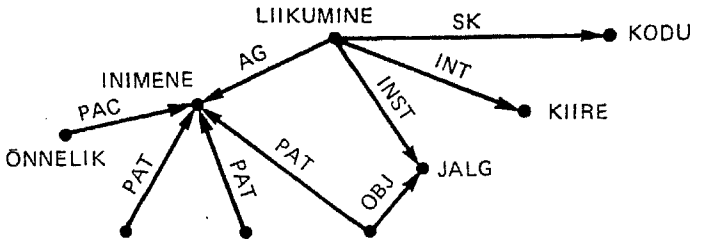
Sõna *jooksma* struktuuris on mõned üksused nurksulgudes. Vastav elementaartähendus ei kuulu otseselt kirjeldatava sõna tähendusse, vaid vaadeldav haru esindab sõna süntaktilist valentsi (selle haru kaudu saab sõna seostuda teiste sõnadega) ning haru tipus olev nurksulgudes elementaartähendus märgib nende sõnade domineerivat semantilist kategooriat, mis saavad esineda antud seose täitja positsioonis. Sõnal *jooksma* on niimoodi märgitud kolm haru: jooksmise agent («kes jookseb»), lähtekoht ja sihtkoht.

Lause süntaktilises esituses, mis on semantilise analüüsi sisendiks, püütakse sõnadevaheliste seostena kasutada võimalust mööda samu sisulisi seoseid nagu semantilistes esitustes. See lihtsustab märgatavalt semantilise analüüsi käiku.

Sõnade asendamist nende semantiliste esitustega on loomulik alustada lause domineerivast üksusest, öeldisest. Tema määrab lõpptulemuseks oleva semantilise struktuuri raamid. Kui mingisse sõnasse suubub süntaktilises esituses seosenool, siis sõna asendamisel tema semantilise esitusega tuleb nool vahetult siduda sõna täienduse domineeriva üksusega.



Joon. 21



Joon. 22

Vaadeldava lause puhul tuleb niisiis kõigepealt asendada sõna *jooksma*. Sõna *poiss* jääb tulemuseks olevas struktuuris täitma agendi (AG) positsiooni, *kodu* sihtkoha (SK) positsiooni; lähtekoha positsioon jääb täitmata. Kui oleme ka need sõnad kirjeldatud viisil asendanud semantiliste esitustega, samuti sõna *õnnelik* (seda sõna, nagu ka sõna *kodu* me oma näites ei analüüsi), saame struktuuri joonisel 22.

Seda võikski lugeda vaadeldava lause semantiliseks esituseks, kui meid ei segaks põhimõttelise tähtsusega asjaolu, millest oli põgusalt juttu eespool lause semantilist esitust kirjeldades. Asi on selles, et keelt mõistva süsteemi puhul ei piisa lause abstraktse tähenduse väljaselgitamisest. Tekst, mida süsteem analüüsib, on talle teadmiste allikaks. Seetõttu peab süsteem lausetest välja lugema konkreetsed faktid, mis puudutavad kindlaid üksikobjekte ja -sündmusi teksti poolt kirjeldatavas maailmas. Antud juhul tähistab näiteks sõna *poiss* ilmselt poissi kui konkreetset isikut, mitte kui abstraktset kategooriat; niisamuti on sõnadega *jooksma* ja *kodu*, samuti sõnaga *õnnelik*. Sellesama poisi, tema kodu, tema kojujooksmise või ka tema õnnelikkuse kohta võib olla tekstis varem midagi öeldud ja võidakse midagi öelda ka järgnevas tekstis. Seetõttu on vaja lause semantilises esituses laskuda üksikobjektide ja -sündmuste tasandile. Tuleb konstrueerida lauses mainitud üksikobjekte, -tegevusi ja -omadusi esitavad üksused (semantilise võrgu tipud) ning siduda need paradigmaatiliste seoste (SUP ja EL) abil vastavate üldiste kategooriatega osutamaks, et tegemist on nende kategooriate üksikeksemplaridega. Teiseks tuleb üksikeksemplare siduvad lauses osutatud süntagmaatilised seosed (AG, PAT ja SK) üldkategooriate tasandilt siia üle kanda. Nii talitades saamegi eespool (joon. 17) esitatud kahekihilise struktuuri.

Kirjeldatu puudutas keelt mõistvate süsteemide semantilise kom-

ponendi ehituse ja töötamise kõige üldisemaid põhimõtteid. Kuid süsteemide semantilise varustuse kirjeldus oleks väga puudulik, kui me sellega piirduksimegi. Protsess, mille tulemusena konstrueeritakse lause semantiline esitus, nii nagu me seda kirjeldasime, ei ammenda kaugeltki teksti mõistmise tegelikku protsessi.

Teksti mõistmine inimese poolt on alati seotud tema teadmistega. Seos on seejuures kahepoolne. Esiteks, tekstist saadud uued andmed lisab inimene oma teadmistele, niisiis muutub mõistmise tulemusena nii või teisiti seniste teadmiste sisu. Teiseks mõjutavad inimese olemasolevad teadmised oluliselt teksti mõistmise protsessi ennast, selle kulgu ja ka tulemust. Tänu eelteadmistele loeb inimene tekstist alati välja rohkem, kui selles vahetult on öeldud. Mida ulatuslikumad on inimese teadmised nähtustest, millest tekstis räägitakse, seda paremini ta teksti mõistab, seda täielikumalt suudab ta mõtestada tekstis esitatud fakte ja väiteid. Mõtestavat interpreteerimist juhtivate mehhanismide uurimine ja nende mehhanismide kas või osalinegi reprodutseerimine arvutis on kujunenud keelt mõistvate süsteemide sõlmprobleemiks.

Peatume siin kahel tekstis sisalduvate andmete täiendamise viisil, mida rakendatakse keelt mõistvates süsteemides. Need on esiteks järelduste tuletamine tekstist ja teiseks teadmisskeemide ehk mäluskeemide kasutamine. Neid saab ainult tinglikult lugeda kaheks erinevaks mehhanismiks, sest teadmisskeemide kasutamine on sisuliselt alati mingite järelduste tuletamine. Kuid nagu allpool võime veenduda, on siiski mõttekas neid kaht teadmistega opereerimise moodust lahus vaadelda.

Järelduste tuletamine kitsas mõttes kui lausete semantiliste esituste teatud loogikalise töötluse protseduur on keelt mõistvates süsteemides tuntud SAD-SAM-ist alates. Esimestes süsteemides kasutati järeldusi siiski väga piiratud ulatuses, 1970-ndatel aastatel muutus see aga keskseks probleemiks. Kujunesid välja arusaamad loomuliku keele mõistmisel vajalike järelduste tüüpidest ja nende rakendamise tingimustest.

Järelduste tuletamise olulisusest teksti mõistmisel rääkisime esimeses peatükis seoses loomuliku keele tekstide omadusega, mida nimetasime semantiliseks elliptilisuseks. Tekst on alati lünklik, selles esitatakse kirjeldatavate sündmuste või nähtuste kohta vaid üksikuid pidepunkte, ülejäänu jääb lugejal või kuulajal juurde mõtelda — järeldada. Inimesel on tekstis öeldu täiendamine omapoolsete järeldustega niivõrd automatiseerunud, et ta lihtsamatel juhtudel üldse ei märkagi, et ta tekstis vahetult öeldule midagi lisab, ning seetõttu ei saa ka aru, et niisugune informatsiooni lisamine võib teksti automaatanalüüsi puhul kujutada endast probleemi. Näiteks lausest *Mees tuli tuppa ja istus laua taha* järeldame täie enesestmõistetavusega, et mees oli enne kuskil väljaspool tuba, pärast tulemist (selle

tulemusena) asus aga toas; ja siit omakorda järeldame, et laud, mille taha mees istus, pidi asetsema samas toas.

Järelduste tuletamise vahendiks on järeldusreeglid. Need on reeglid kujul A→B («kui A, siis järelikult ka B»), mis ütlevad, et kui teadmistes sisaldub struktuur A, võime lisada teadmistesse ka struktuuri B.

Esimene olulisem järeldusreeglite liik on seotud elementaartähendustega. Iga elementaartähendust iseloomustab kindel hulk järeldusreegleid, mis võimaldavad järeldada midagi antud elementaartähendusest lähtudes. Need reeglid tegelikult väljendavadki elementaartähenduse sisu automatiseeritud süsteemi jaoks, sest nad määravad, missugused tagajärjed on süsteemi jaoks antud elementaartähenduse esinemisel vastuvõetavas tekstis.

Järeldusreeglitele teksti mõistmise vahendina on erilist tähelepanu pööranud R. Schank. 1970-ndate aastate alguses tema juhtimisel loodud süsteemis MARGIE on järeldusreeglid üks peamisi keele mõistmist modelleerivaid mehhanisme. Eespool tõime Schanki elementaartegude loendi. Igaühega neist on Schankil seotud järeldusreeglid, mida süsteem vastavat elementaartähendust kohates rakendab. Näiteks elementaarteoga A PTRANS B kohalt C kohta D «A liigutab objekti B kohalt C kohale D» on seotud järgmised olulisemad järeldused:

1) B asetses varem kohal C;
2) B asetseb pärast PTRANS-i kohal D;
3) B ei asetse enam kohal C.

Elementaarteost A INGEST B «A sööb B» järelduvad aga järgmised faktid:

1) PTRANS kui tegevuse INGEST teostamise vahend (ja selle kaudu kõik PTRANS-i järeldused);
2) objekt B lakkab eksisteerimast esialgsel kujul;
3) Kui B on söödav, on A «toitumus» INGEST järel «parem»; jne.

Kui järeldusreeglid on niimoodi elementaartähenduste juures kord formuleeritud, on nad rakendatavad põhimõtteliselt kõikidele sõnadele, kus vastavad elementaartähendused esinevad. Näiteks PTRANS-iga seotud järeldused on seeläbi tuletatavad kõigist sõnadest, mis tähistavad ruumilise liikumise mingit vormi (*minema, tulema, lendama, veerema* jne.; vrd. eeltoodud näitelauset sõnaga *tulema*). Et elementaartähendusi on märksa vähem kui sõnu, võimaldab niisugune moodus ökonoomselt kirjeldada paljusid sõnade semantilisi omadusi.

Elementaartähendustega seotud järeldusreeglid pole siiski ainsad reeglid, millele tugineb keelt mõistva süsteemi järeldamisvõime. Kindlate elementaartähendustest moodustatud struktuuridega, sealhulgas näiteks sõnadega võivad olla seotud omad järeldusreeglid, mis ei tulene ühestki antud struktuuri

kuuluvast elementaartähendusest eraldi võetuna. Niisugused reeglid väljendavad eelkõige mitmesuguseid sisulisi seaduspärasusi, mis kehtivad analüüsitava teksti ainevallas. Nende reeglite näitena võib tuua mõned reeglid, mis on vajalikud inimeste eesmärgipäraseid tegevusi kirjeldavate tekstide mõistmiseks:

1) *A teeb tahtlikult B → A püüab tegevuse B abil esile kutsuda mingit olukorda C (A omab eesmärki C);*
2) *A teeb B eesmärgiga saavutada C → A eeldab, et ilma tema tegevuseta B olukord C ei teki.*

Edasi võib järeldusreeglite hulgas nende mõju suuna järgi esile tuua kaks olulist rühma: ettepoole suunatud ja tahapoole suunatud järeldused. Esimesed osutavad sündmustele või olukordadele, mis hakkavad kehtima pärast vahetult analüüsitavat sündmust. Teised osutavad sündmustele või olukordadele, mis pidid aset leidma enne antud sündmust. Esimesed reeglid toovad esile tegevuste tagajärjed, teised fikseerivad nende eeltingimusi ja põhjusi. Teksti mõistmise seisukohalt on need järelduste liigid olulised seepärast, et võimaldavad analüüsitavat lauset siduda eelneva ja järgneva tekstiga. Kõnealused kaks järeldusreeglite tüüpi on eristatavad nii elementaartähenduste kui ka komplekssemate semantiliste struktuuride tasandil. Näiteks üksuse PTRANS eespool toodud järeldustest on esimene tahapoole (minevikku) suunatud järeldus, teised kaks ettepoole (tulevikku) suunatud järeldused. Schanki süsteemis on lisaks eespool tutvustatud elementaartegudele kasutusel ka kindel hulk elementaartähendusi tähistamaks olukordi, seisundeid ja objektidevahelisi vahekordi, nende hulgas näiteks PAIKNEMINE (objekti ruumiline asetus teiste objektide suhtes), KONTROLL (kontrolli omamine teatud objektide või olendite üle), TEADMINE (informatsiooni omamine mälus) jt. Elementaarteod ja seisundid on vastastikku seotud reeglitega, mis osutavad, missugustes seisundites on võimalikud missugused teod ning missugused seisundid võivad olla missuguste tegude tulemuseks. Näiteks järgmistes seisundites on võimalikud vaid järgmised teod (s. t. tegudest järelduvad vastavad seisundid kui nende eeltingimused):

PAIKNEMINE ⟷ PTRANS
KONTROLL ⟷ ATRANS
INGEST
TEADMINE ⟷ ATRANS
MBUILD

Seni kirjeldatud järeldusreeglid on üldkehtiva, loogikalise iseloomuga selles mõttes, et niipea, kui kehtib reegli vasakul poolel esitatud struktuur, peab paratamatult kehtima ka paremal pool seisev struktuur, seost nende kahe struktuuri vahel ei saa mõjutada mingid empiirilised asjaolud. Niipea, kui näiteks kehtib PTRANS, peavad vältimatult kehtima ka faktid, mis on ülal fikseeritud tema järel-

dustena. Kuid inimesed ei opereeri teksti mõistmisel ainult niisuguste rangelt loogikaliste, sajaprotsendiliselt kehtivate järeldustega. Teksti interpreteerimisel moodustavad väga suure osa järeldustest tõenäosuslikud järeldused, mille põhimõtteline struktuur on: kui kehtib A, siis harilikult kehtib ka B. Nende rakendamise põhimõte seisneb järgmises: kui tekstis on väidetud A, loetakse kehtivaks ka B, kui tekstis pole otseselt öeldud vastupidist. Kõnealused järeldused kajastavad eelkõige sündmuste ootuspärast kulgu, nende tüüpilisi tagajärgi ja põhjusi. Seetõttu kehtivad need järeldusreeglid valdavalt elementaartähenduste tasandist kõrgemal, komplekssete sündmusstruktuuride tasandil. Näiteks kehtib ilmselt järeldusreegel: *kui A käsib B-l teha C, siis ootuspäraselt B ka teeb C.* Selle reegli põhjal järeldame lausest *Direktor käskis sekretäril kaustad kappi panna ja tõttas ise nõupidamisele*, et sekretär pani kaustad kappi — kui järgnevas tekstis otseselt ei öelda, et sekretär seda ei teinud.

Et järeldused loetavatest või kuuldavatest lausetest moodustavad olulise osa teksti mõistmisest, selles pole ilmselt kahtlust. Seetõttu kujunes 1970-ndate aastate alguses isegi arvamus, et keelt mõistev süsteem on seda parem, mida rohkem järeldusi ta suudab mingist lausest tuletada. Kuid see, mis niimoodi abstraktselt sõnastatuna näib positiivse joonena, osutub konkreetses kontekstis konkreetsele lausele rakendatult hoopis negatiivseks nähtuseks, mille vältimiseks on tarvis keelt mõistev süsteem varustada eriliste vahenditega.

Probleem tekib nimelt seotud teksti analüüsimisel. On hästi teada, et sellise teksti tähendus ei ole temasse kuuluvate lausete tähenduste lihtne kogum, vaid tervik, mis kujuneb lausete tähenduste sidumisel kindlate seostega. Ent lausetevahelised seosed on harva otseselt tekstis osutatud, enamasti tuleb need järeldada. Kohates näiteks tekstis kaht järjestikust lauset *Ta ei tule täna tööle. Ta on haige,* tõlgendame neid nii, et teine lause esitab esimeses lauses mainitud sündmuse põhjuse.

Püüdes modelleerida seotud teksti mõistmist, asetaski näiteks Schank süsteemi MARGIE esialgses variandis peamise rõhu järelduste tuletamisele. Iga järjekordse lause puhul teostati semantiline analüüs ning selgitati lause semantiline esitus (elementaartähenduste semantiline võrk). Seejärel püüti sellest tuletada kõik võimalikud järeldused, rakendades järjest kõiki reegleid, siis järeldused neist järeldatud struktuuridest jne. Kui niisuguses järelduste tuletamise protsessis mõni järeldustest haakus kindlal viisil mõne eelneva või järgneva lausega (või mõnega neist lausetest tuletatud järeldustest), loetigi seos lausete vahel sellega kehtestatuks.

Eespool toodud lausepaari puhul võib seda protsessi kujutleda järgmiselt. Analüüsides lauset *Ta ei tule täna tööle* ning püüdes tuletada sellest kõiki võimalikke järeldusi, teeb süsteem mingid järeldu-

sed ka antud fakti võimalike põhjuste kohta. Üheks võimalikuks põhjuseks, miks keegi inimene mingit konkreetset tegevust ei soorita, on see, et tal pole võimalik seda teha; selle mittevõimalikkuse üks võimalikke avaldusvorme on see, et inimene pole sellises füüsilises seisundis, mida tegevus eeldab. Analüüsides seejärel lauset *Ta on haige,* teeb süsteem kindlaks, et selles väidetu sobibki eelmisest lausest järeldatud «nõutava füüsilise seisundi puudumise» kui tööle mittetulemise põhjuse konkreetseks realisatsiooniks, ja kehtestab kahe lause semantiliste esituste vahel vastava seose.

Tegelikkuses pole kõigi võimalike järelduste tuletamise printsiip rakendatav, sest vähegi ulatuslikuma järelduste tuletamise võime puhul on igast lausest tuletatavate järelduste, järelduste järelduste jne. hulk tohutu. Nagu selle printsiibi kriitikud irooniliselt märkisid, jõuaks niisugune süsteem, analüüsides lauset *John tappis Mary,* kunagi kindlasti ka järelduseni, et selle tapmise üheks põhjuseks oli asjaolu, et John ja Mary kunagi sündisid, neil sündimistel aga olid omad põhjused, ja nii edasi, kuni maailma alguseni välja. Ülal esitatud arutluskäik kahe lause vahelise seose väljaselgitamiseks võis tunduda arukas, kuid seda üksnes seetõttu, et mainimata jäid kõik ülejäänud, antud seose seisukohalt mitterelevantsed järeldused.

Ehkki inimestel mängivad järeldused teksti mõistmisel olulist rolli, on päris selge, et järelduste tuletamisel ei toimi nad huupi ega võta läbi kõiki võimalikke järeldusi. Iga lause puhul teeb inimene üksnes järeldused, mis on vajalikud antud kontekstis, mõistmaks analüüsitavat lauset ja selle seost teiste lausetega. Väga suurt osa loogikaliselt võimalikest järeldustest inimene ei vaatle, ta lihtsalt ei näegi neid. Ühest ja samast lausest tehakse erinevates kontekstides erinevad järeldused, mistõttu ka samakujulise lause tähendus (interpretatsioon) kujuneb erinevates kontekstides erinevaks. Näiteks lause *John jooksis silla alla* tähendust tõlgendame järgmistes lausepaarides hoopis erinevalt: 1) *Äkki hakkas vihma valama. John jooksis silla alla;* 2) *Eemalt hakkas kostma politseiauto sireen. John jooksis silla alla.*

Inimesel on niisiis lisaks järelduste tuletamise mehhanismile veel mingid mehhanismid, mis suunavad järelduste tuletamist ja hoiavad seda protsessi alati kindlates piirides. Kui kirjeldatud nähtusesse pisut rohkem süveneda, veendume, et nende suunavate mehhanismide aluseks on inimese t e a d m i s e d situatsioonist, mida tekst käsitleb. Toodud lausepaarides määrab esimene lause üldise situatsiooni, mille kontekstis me tõlgendame teist lauset, ja et situatsioonid on erinevad, on ka tõlgendused erinevad. Seega on keelt mõistvas süsteemis, mille eesmärgiks on modelleerida inimlikku mõistmist, vaja ka inimteadmiste mudeleid, mis kajastaksid teadmiste organisatsiooni inimese mälus ja oleksid kasutatavad teksti mõistmist juhtivate struktuuridena.

Niisuguste struktuuride kirjeldamiseks võeti kasutusele teadmisskeemi mõiste, mille aluseks on järgmine hüpotees. Inimese — ja ilmselt mitte ainult inimese — teadmiste (kogemuse, mälu) üheks põhiliseks korraldusvormiks on teadmiste säilitamine teatud üldistatud nähtuste, situatsioonide ja tegevuste järgi. Inimese mälus on andmed suure hulga stereotüüpide — sageli korduvate, ehkki detailides varieeruvate nähtuste kohta, mida ta tunneb oma isiklikust kogemusest või mingist muust allikast: «sünnipäevapidu», «jalgpallimatš», «restoranis», «teatris», «perekonnatüli», «vestlus ülemusega», «vestlus vana tuttavaga, keda juhuslikult kohatakse» jne. Üldistatud ja tüpiseeritud teadmised niisuguste nähtuste kohta moodustavadki teadmisskeemi. Teadmisskeemid võivad olla väga erineva üldistusastmega, alates skeemidest nagu «inimese elukäik» ja lõpetades skeemidega nagu «kingapaelte kinnisidumine» või «õllepudeli avamine». Skeemide hulk keskmise inimese mälus peab olema tohutu. Inimene püüab oma igapäevases käitumises kõikjal näha tüüpilist, leida igas uues situatsioonis juba tuttavate tüüpsituatsioonide elemente. See on talle vajalik selleks, et osata uues situatsioonis käituda.

Samasugune olukord kehtib teksti analüüsimisel. Kõnealuse hüpoteesi järgi seisneb oluline osa teksti mõistmisest selles, et lugeja viib tekstis kirjeldatu mingi teadmisskeemi alla ja sellest lähtudes tõlgendab teksti osi (kuni üksiklausete ja nende osadeni välja) kui struktuure, mis määratlevad selle skeemi teatud aspekte. Selle asemel, et tuletada lausetest kõik võimalikud järeldused ja otsida huupi nende seoseid eelnevate või järgnevate lausetega, võib nüüd kõigepealt piirduda järeldustega, mis on vajalikud skeemi mingite aspektide täitmiseks konkreetse materjaliga. Seda esiteks. Teiseks aitavad teadmisskeemid teksti mõistmisele kaasa sel teel, et nad lisavad informatsiooni, mida tekstis ei teatata. Tekstis jäetakse mainimata eelkõige need faktid ja seosed, mida loetakse küllalt harilikeks, tüüpilisteks ja seetõttu vastavas olukorras enesestmõistetavateks. Eeldatakse, et inimesel on kirjeldatava nähtuse teadmisskeem, mis sisaldab vastavaid andmeid ja lisab puuduvad lülid. Kujutlegem näiteks, kui palju taipab korvpalli- või jäähokireportaažist inimene, kel pole neist mängudest mingit ettekujutust, võrreldes inimesega, kes üksipulgi teab mängureegleid.

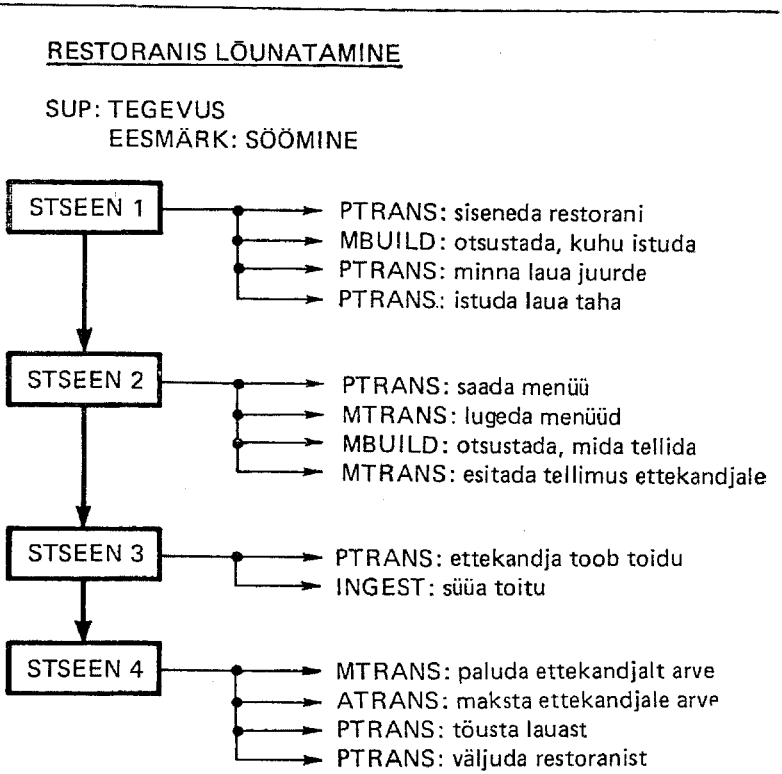
Teadmisskeemi mõiste tõi tehisintellektialastesse käsitlustesse M. Minsky 1970-ndate aastate keskel.[45] Minsky ise nimetas neid teadmiste struktuure «raamideks» (ingl. k. *frame*). Teadmisskeemid kujutavad raame selles mõttes, et nad määravad kindlaks vastavate situatsioonide või sündmuste põhielemendid, piirjooned, ja jätavad vabaks teatud tühemikud, mis täi-

[45] M. Minsky selleteemalised põhilised tööd on tõlgitud ka vene keelde, näit. М. Минский. Фреймы для представления знаний. М., 1979.

detakse konkreetse materjaliga, näiteks tekstist saadud andmetega situatsiooni või sündmuse kohta. Pärast Minsky tööde ilmumist algas mitte ainult keelt mõistvate süsteemide loomises, vaid tehisintellektis tervikuna «teadmiste ajastu», mis avaldub eelkõige selles, et on välja töötatud teadmiste ning nende kasutamismeetodite väga mitmesuguseid käsitlusi. Tähelepanu keskpunkti tõusid seejuures, seda eriti keelt mõistvate süsteemide puhul, nn. argiteadmised *(common sense knowledge)*, mis on inimeste igapäevase käitumise aluseks. Just seda tüüpi teadmised, nagu «Kui inimesed jäävad vihma kätte, siis saavad nad märjaks» ja «Harilikult inimestele ei meeldi märjaks saada» jäävad tekstides kõige sagedamini väljendamata, ühtlasi aga jäävad kõige kergemini kahe silma vahele uurijatel, kes püüavad konstrueerida teksti mõistmise protsessi formaalset mudelit.

Formaalselt võib teadmisskeemi kujutada hierarhilise struktuurina, milles kõrgemate tasandite elemendid esitavad väiteid, mis peavad kehtima igas antud skeemi poolt kirjeldatavas situatsioonis, madalamate tasandite elemendid esitavad aga muutujaid, mis erinevates konkreetsetes situatsioonides võivad olla täidetud mitmesuguse antud muutuja poolt esitatud nõudeid rahuldava materjaliga. Teadmisskeemi elementideks võivad olla teised, spetsiifilisemaid tegevusi või olukordi kajastavad teadmissituatsioonid. Näiteks teadmisskeem «restoranis lõunatamine» sisaldab ilmselt elemente «sisenemine restorani», «tellimuse esitamine», «söömine», «arve tasumine» jne., millest igaüks on omakorda teadmisskeem oma kohustuslike komponentidega ja varieerimisvõimalustega.

Teadmisskeemide ideed on rakendatud väga erinevates vormides. Eri nähtustüüpide jaoks on vaja teatud määral erinevaid teadmisskeeme. Näiteks teadmised tegevustest või protsessidest on teistsuguse ehitusega kui teadmised esemetest või staatilistest olukordadest. Toome näitena joonisel 23 R. Schanki teadmisskeemi «Restoranis lõunatamine» (Schank ise nimetab selliseid skeeme stsenaariumideks). Esitus on mõeldud vaid illustratsioonina ja kujutab teadmisskeemi põhimõttelist ehitust, mitte selle täpset vormi. Kõigepealt on osutatud teadmisskeemi ülemmõiste (SUP), s. o. mõistete klass, kuhu kirjeldatav mõiste kuulub — antud juhul tegevus, mille eesmärk on söömine; ka sellele üldisimale mõistele vastab reeglina oma teadmisskeem. Üksikute stseenide vahel kehtib ajaline järjestus. Stseenide sisu on kirjeldatud elementaartähenduste abil, näites on esitatud iga üksiksündmuse puhul vaid selle keskne elementaartegu. Sündmused on stseenis seotud kindlal viisil: iga teo tulemuseks on olukord, kus on võimalik järgmine tegu. Keerulisemates teadmisskeemides-stsenaariumides võib sisalduda mitmeid tegevuse kulgemise rööpliine, mille valik sõltub kindlatest tingimustest; võib olla osutatud, missugused on vastava tegevuse kohustuslikud ja mis-



Joon. 23

sugused vaid võimalikud komponendid.

Kui tekst kirjeldab tegevust, mille jaoks on olemas teadmisskeem — nagu antud juhul restoranis söömas käimine —, konstrueerib seda analüüsiv süsteem vastava teadmisskeemi konkreetse eksemplari, kus püüab tema abstraktse struktuuri täita andmetega, mis leiduvad otseselt tekstis või on sellest järeldatavad. Kehtivaks loetakse ka väited, mis sisalduvad teadmisskeemis, kuid mille kohta tekstis ei sisaldu otseseid andmeid. Näiteks võib restoranis lõunatamise kohta tekstis olla vaid lause *Jüri lõunatas «Kaunases»*. Kuid teadmisskeemi põhjal võib süsteem vastata ka näiteks küsimustele, kas Jüri «Kaunases» istus, kas ta seal kellelegi raha maksis jms. Niisamuti annab teadmisskeem vajalikud andmed lausetevaheliste seoste

määramiseks, andmed «puuduvate lülide» kohta sündmuste kirjeldustes. Kohates tekstis näiteks lauseid *Jüri oli «Kaunases» lõunat söömas. Kui ta arvet maksma hakkas, avastas ta, et rahakott on kadunud,* ei ole süsteemil tarvis hakata lausetest juhuslikus järjekorras järeldusi tuletama, selleks et kindlaks teha, kuidas on seotud restoranis lõunatamine ja arve maksmine.

Teadmisskeemide käsitluse lõpetuseks märgime, et asja tuum pole loomulikult neis konkreetsetes teadmistes, mida ühe või teise valdkonna tekstide mõistmiseks teadmisskeemide ja teiste teadmisstruktuuride kujul esitatakse. Keelt mõistvate süsteemide loomisel on põhiliseks huvi- ja diskussiooniobjektiks kujunenud inimteadmiste organisatsiooni üldised põhimõtted. Sellel alusel on märgatavalt tihenenud ka lingvistide ja psühholoogide koostöö, mille näiteks on kas või R. Schanki ja psühholoog R. P. Abelsoni raamat «Stsenaariumid, plaanid, eesmärgid ja mõistmine. Uurimus inimteadmiste struktuurist».[46] Selles raamatus väidetakse: uurimuse eesmärk pole mitte keelealased teadmised traditsioonilises mõttes, s. o. morfoloogilised, süntaktilised ja semantilised teadmised, vaid inimeste üldised teadmised, mida nad kasutavad, tuletades tekstist järeldusi ja tõlgitsedes lausejärjendeid kui terviklikke jutustusi.

Võtame nüüd lõpuks lühidalt kokku, missugustest osadest koosneb tüüpiline keelt mõistev süsteem. Süsteemi keelevara kitsas mõttes moodustavad sõnastik ja analüüsi- ning sünteesireeglite plokk, millesse kuuluvad nii morfoloogilise, süntaktilise kui ka semantilise analüüsi ja sünteesi reeglid. Reeglite plokk ja sõnastik kattuvad osaliselt, sest nagu nägime, on ka mitmed sõnastikus esitatud andmed tõlgendatavad reeglitena. Süsteemi teine põhikomponent on mälu, mis samuti koosneb kahest põhimõtteliselt erinevast osast. Ühe osa moodustavad süsteemi üldised teadmised; sellesse kuuluvadki järeldusreeglid ja teadmisskeemid, millest rääkisime. Seda mäluosa on sageli nimetatud semantiliseks mäluks. Teises mäluosas sisalduvad süsteemi teadmised konkreetsetest objektidest, sündmustest ja faktidest, mis ta on ammutanud analüüsitud tekstidest; see on andmemälu või ka episoodiline mälu. Terminid «semantiline mälu» ja «episoodiline mälu» on üle võetud psühholoogiast. Nende erinevust võib kujundlikult iseloomustada nii: episoodiline mälu on nagu päevik, kus on kirjas kõik inimese poolt läbielatud sündmused; semantiline mälu on nagu entsüklopeedia, mis on saadud päeviku andmete üldistamise ja süstematiseerimise tulemusena.

Peatüki lõpuosa pühendame tuntumate keelt mõistvate süsteemide kirjeldamisele.

[46] Schank R. C., Abelson R. P. Scripts, plans, goals and understanding. An inquiry into human knowledge structures. Hillsdale, 1977.

4.4. Keelt mõistvaid süsteeme 1970-ndatest aastatest

1970-ndate aastate alguse tuntumaid süsteeme on kindlasti T. Winogradi poolt Massachusettsi Tehnoloogiainstituudis loodud keelt mõistev «robotimudel» SHRDLU. See süsteem imiteeris roboti tegevust, millele sai loomulikus (inglise) keeles anda ülesandeid ning esitada mitmesuguseid küsimusi tema «maailma» ja tema tegevuste kohta selles maailmas. SHRDLU maailm oli võrdlemisi rangelt piiratud: selleks oli laud, millel asusid kast ja mitmesuguse kuju, suuruse ning värvusega esemed — risttahukad, kuubikud ja püramiidid. Nende esemetega võis SHRDLU manipuleerida vastavalt käskudele; manipuleerimist sai jälgida kuvaril.

Esitame tüüpilise dialoogi inimese (I) ja SHRDLU (S) vahel.

1. I: *Tõsta üles suur punane risttahukas.*
   S: *OK.*

Süsteem vastab «OK», kui ta on käsust aru saanud ja täidab selle. Antud juhul tähendab käsu täitmine seda, et robot viib käe nimetatud eseme juurde ja tõstab selle üles; kui eseme peal oli mõni teine ese, siis tõstab robot selle enne kõrvale.

2. I: *Võta laualt püramiid.*
   S: *Ma ei tea, missugust püramiidi sa mõtled.*
   Laual on parajasti kolm püramiidi.
3. I: *Leia üks risttahukas, mis on suurem kui see, mida sa käes hoiad, ja tõsta ta kasti.*
   S: *Ma eeldan, et ta all mõtled sa risttahukat, mis on suurem kui see, mida ma käes hoian. OK.*

Näide demonstreerib, et SHRDLU saab edukalt hakkama asesõnadega *mis, see, ta* tähistatud objektide õige valikuga.

4. I: *Mis on kastis?*
   S: *Sinine püramiid ja sinine risttahukas.*
   I: *Millele püramiid toetub?*
   S: *Kasti põhjale.*

Antud juhul, erinevalt küsimusest nr. 2, ei teki süsteemil arusaamatust selles suhtes, missugust püramiidi silmas peetakse: süsteem eeldab, et mõeldud on püramiidi, mida ta eelmises vastuses mainis.

5. I: *Kas sa olid liigutanud mõnd püramiidi, enne kui asetasid rohelise püramiidi väikesele kuubikule?*
   S: *Jah, rohelist.*
   I: *Millal sa teda liigutasid?*
   S: *Siis, kui ma tõstsin üksteise peale punase kuubiku, suure punase risttahuka ja rohelise kuubiku.*
   I: *Milleks sa teda liigutasid?*
   S: *Et teda kõrvale tõsta.*
   I: *Ja milleks sa seda tegid?*
   S: *Et punast kuubikut pealt tühjaks teha.*

T. Winogradi süsteemi kõige enam muljet avaldav omadus oligi

tema võime mõista ja kasutada keerulisi keelekonstruktsioone, mis sisaldasid asesõnu, kõrvallauseid ja elliptilisi väljendeid. Seejuures on selge, et tegemist polnud ladusa vestluse palja imitatsiooniga nagu ELIZA puhul. SHRDLU vastused näitavad, et ta orienteerus «risttahukate maailmas» ja oskas ka seletada ning kommenteerida enese tegutsemist selles. SHRDLU vastused on alati vastused just sellele, mida küsitakse.

Winogradi süsteemi lingvistiliseks aluseks oli inglise keeleteadlase M. A. K. Halliday keeleteooria — süsteemne grammatika. Erilised olid ka süsteemi semantilise analüüsi põhimõtted. Nii sõnade kui ka lausete tähendusi tõlgendati kui protseduure. Eriti küsimusi ja käske püüti interpreteerida kui protseduure, mis viiksid vastuse konstrueerimisele või nõutud tegevuse sooritamisele. Niisugune «protseduraalne semantika» osutus tegutseva robotiga suhtlemisel igati tõhusaks. Kuid ühtekokku olid nii SHRDLU keelelise analüüsi kui ka semantiliste mudelite põhimõtted niivõrd tihedalt seotud selle konkreetse «minimaailma» seaduspärasustega, milles SHRDLU tegutses, et ei Winograd ise ega ka keegi teine pole hiljem püüdnud neidsamu analüüsimeetodeid laiendada muudele aladele.

Teine 70-ndate aastate alguspoole tuntum süsteem oli R. Schanki juhtimisel loodud MARGIE, millest eelmises jaotises oli siin-seal juttu. Nagu seal osutatud, huvitasid Schanki MARGIE loomise ajal ja huvitavad ka praegu esmajoones keele mõistmise teoreetilised probleemid. MARGIE oli eeskätt Schanki teoreetilise kontseptsiooni illustratsiooniks.

Schanki esimene tees oli, et teksti mõistmise tulemusena saadud teadmised tuleb säilitada universaalses, keelest sõltumatus esitusviisis, omamoodi interlingvas. Selleks töötas ta välja oma kontseptuaalse sõltuvusesituse, mille universaalsuse rõhutamiseks vältis keelest võetud sõnade kasutamist selle üksuste tähistajatena (vt. eelmises jaotises toodud elementaartegude loendit).

Nagu osutatud, rõhutati MARGIE-s samuti järelduste tuletamist teksti mõistmise olulisima komponendina. MARGIE töötaski kahel viisil: parafraseerides talle esitatud lauseid ja tuletades lausetest järeldusi.

Järelduste tuletamiseks töötati välja terve järelduste tüpoloogia, eristades järgmisi järelduste põhitüüpe:

1) järeldused nähtuste põhjuste kohta *(John lõi Maryt → John oli vihane Mary peale);*

2) järeldused tagajärgede kohta *(John lõi Maryt → Mary sai haiget);*

3) järeldused kaasnevate, s. o. samaaegselt kehtivateks loetavate nähtuste kohta *(John ütles Maryle, et Bill külastas Ritat → John teadis, et Bill külastas Ritat);*

4) järeldused, mis kajastavad subjektide oletatavaid tegevusi vastavas situatsioonis *(John on otsustanud osta uued kingad → John läheb ilmselt kingapoodi);*

5) järeldused nähtuste normaalsete seoste kohta *(Kus on Sam kell üheksa hommikul? → ilmselt tööl)*.

Igaüks neist põhitüüpidest jagunes reaks alltüüpideks vastavalt reeglites osalevate objektide ja nähtuste konkreetsematele tüüpidele.

MARGIE, nii nagu ka Winogradi SHRDLU kuulusid veel perioodi enne teadmiste esitamise ja kasutamise meetodite muutumist keelt mõistvate süsteemide võtmeprobleemiks. Need viimased mõisted on kesksteks Schanki järgmises süsteemis SAM *(Script Applier Mechanism)*. Üksiklausete mõistmise, nende parafraseerimise ja järelduste tuletamise modelleerimiselt asuti piiratud temaatikale pühendatud seotud tekstide mõistmise modelleerimisele. Põhiline SAM-i teadmiste korraldusvorm on stsenaariumid, mis on teadmisskeemide erikujud. Peamiseks teemaks on millegipärast valitud restoranide külastamine, millele vastava stsenaariumi üldist struktuuri kirjeldasime eelmises jaotises. SAM kasutab seda stsenaariumi ja temaga seotud mitmesuguseid teisi stsenaariume selleks, et ümber jutustada ja resümeerida restoraniteemalisi jutukesi või vastata nende kohta esitatud küsimustele. Järgmine jutuke on tüüpiline näide neist, mida SAM analüüsib:

John läks restorani. Hostess juhatas Johni kohale. Hostess tõi Johnile menüü. Kelner tuli laua juurde. John tellis homaare. Johni teenindati kiiresti. John andis kelnerile rikkaliku jootraha. John lahkus restoranist.

Selle jutu põhjal vastab SAM näiteks järgmistele küsimustele:

K(üsimus): Milleks läks John restorani?
V(astus): Selleks et seal süüa.
K: Kas kelner tõi Johnile menüü?
V: Ei, menüü tõi hostess.
K: Milleks hostess andis Johnile menüü?
V: Et John oskaks tellida.
K: Mis sai edasi, kui hostess oli andnud Johnile menüü?
V: John luges menüüd. Kelner nägi, et John istub lauas. Kelner läks laua juurde.
K: Mida John sõi?
V: Homaare.
K: Kes tõi Johnile homaarid?
V: Kelner.
K: Mis sai siis, kui John oli homaarid ära söönud?
V: John andis kelnerile märku, et ta tahab arvet tasuda. Kui kelner tõi arve, andis John talle rikkaliku jootraha.
K: Kas homaarid meeldisid Johnile?
V: Tõenäoliselt.
K: Kes maksis arve?
V: John.

Lugeja võib ise otsustada, missugustele küsimustele vastamiseks piisab otseselt tekstis toodud andmetest ja missuguste küsimuste puhul kasutab SAM üldisi teadmisi restoranis käimisest.

Lisaks stereotüüpseid situatsioone või tegevusi kirjeldavatele stsenaariumidele kasutab SAM ka «isiksusestsenaariume» — teadmisskeeme inimtüüpidest ja nende käitumisest standardsituatsioonides.

Neid stsenaariume läheb vaja juttudes, kus tegelastel ilmnevad kindlad iseloomujooned.

Schanki tähelepanu on koondunud just inimtegevusi kirjeldavate tekstide mõistmisele. Lisaks stsenaariumidele kasutab ta inimtegevuste kirjeldamiseks ka teisi, stsenaariumidest põhimõtteliselt erinevaid teadmisstruktuure, millest olulisimad on p l a a n i d ja t e e m a d.

Stsenaariumid kui teadmisskeemid juhivad inimeste käitumist stereotüüpsetes situatsioonides. On aga küllalt olukordi, kus ükski standardne tegevusjärjend ei ole täielikult kasutatav ning inimesel tuleb konstrueerida eri plaan. Et mõista teksti, kus kirjeldatakse inimese tegevust niisuguses mittestandardses olukorras, tuleb eelkõige aru saada inimese plaanist, seetõttu on teadmised plaanidest vajalikud ka keelt mõistvatele süsteemidele.

Plaani koostamiseks kasutatakse Schanki süsteemis kaht liiki teadmisstruktuure: «delta-eesmärke» ja «plaanifragmente». Delta-eesmärkide all mõeldakse üldistatud, kuid küllalt tihti alleesmärkidena esinevaid situatsioonimuutusi. Näiteks Δ PROX(x, y) on muutus, mille tulemusena hakkab kehtima PROX (x, y), s. o. «x asetseb vahetult y juures». Iga delta-eesmärgiga on seotud rida plaanifragmente — alternatiivseid tegevusskeeme, mis võivad viia vastava eesmärgini. Plaani koostamine seisnebki lõppeesmärgist lähtudes vahepealsete eesmärkide formuleerimises delta-eesmärkide termineis ja iga delta-eesmärgi puhul olukorrale vastava plaanifragmendi valikus. Plaan koostatakse põhimõtteliselt iga kord uuesti ja pärast kasutamist mõne aja möödudes «ununeb». Kuid kui mõni plaan on osutunud tervikuna kasutatavaks mitmel korral, võib see omandada stsenaariumi staatuse, see tähendab, ta säilitatakse vastaval kujul süsteemi mälus.

«Teemad» kujutavad endast kõige üldisema taseme teadmisi inimestest, nende käitumisest ja eesmärkidest. Need on näiteks teadmised, kuidas tüüpiliselt käitub armunud inimene või mida tähendab olla ühe või teise ameti või elukutse esindaja (osariigi kuberner, politseinik, kelner), kuni teadmisteni sellest, missugused on inimeste tüüpilised eesmärgid elus: elada luksuses, võidelda õigluse eest, saada kuulsaks, saada arstiks.

Eriti viimased, teemade tüüpi teadmised illustreerivad rohkem Schanki ja ta kolleegide teoreetilisi arusaamu sellest, missuguseid teadmisi peaks omama tõeline keelt mõistev süsteem, kui seda, mida süsteem SAM tegelikult teab ja oskab. Niisuguste globaalsete teadmiste realiseerimine arvutitel on tänapäeval võimalik vaid võrdlemisi piiratud ulatuses.

Schanki juhtimisel Yale'i Ülikoolis tehtud tööst on võrsunud mitmed projektid, mis jätkavad sama põhimõttelist suunda ka pärast 1980. aastat.

R. Wilensky arendab California Ülikoolis Berkeleys kaht süsteemi, PAM *(Plan Applier Mechanism)* ja PHRAN *(PHRasal ANalyzer)*.

Need süsteemid kasutavad ühelt poolt teadmisi inimeste tegevustest, eesmärkidest ja arutlusviisidest, teiselt poolt keelelisi teadmisi tegevusi kirjeldavate tekstide (juttude) ülesehitusest, selleks et mõista juttude sisu.

J. Kolodner, kes jätkab Yale'is alustatud tööd Georgia Tehnoloogiainstituudis, on keskendunud eelkõige teadmiste säilitamise viisidele ja seaduspärasustele, mis on vajalikud ulatuslikumaks tekstide mõistmiseks. Tema süsteemi CYRUS «andmebaasiks» on episoodid omaaegsete USA riigisekretäride Cyrus Vance'i ja Edmund Muskie tegevusest.

Tehisintellekti alal on eriti hästi tuntud Stanfordi Rahvusvahelise Uurimisinstituudi *(SRI International)* tehisintellektikeskuse tööd. Siinsed viimaste aastate uuringud loomuliku keele mõistmiseks on niisamuti keskendunud kahele põhisuunale: keeles (tekstis) eneses toimivate seaduspärasuste detailsemale ja sügavamale mõistmisele ning inimteadmiste ja -arutluste modelleerimisele.

Esimest suunda esindab projekt TEAM, mis hõlmab mitut allprojekti ja mille raames arendatakse tõhusamaid süntaktilise ja semantilise analüüsi meetodeid, muuhulgas selliseid, mida võiks kasutada praktilisel suhtlemisel andmebaasidega ning mis seejuures oleksid sobitatavad eri ainevaldkondi esindavate ja eri struktuuriga andmebaasidega.

Teist suunda esindab projektide pere KLAUS *(Knowledge Learning And Using System)*, mille eesmärk on arendada fundamentaalseid teoreetilisi uuringuid teadmiste ja arutlusmehhanismide, semantika ja pragmaatika valdkonnas.

Lõpetuseks

Läbi kolme peatüki oli juttu aladest, kus tänapäeval kasutatakse arvuteid loomuliku keele tekstide analüüsimiseks. Loodame, et lugejale sai selgemaks, kuidas võib arvutitele keele õpetamine aidata üle saada probleemidest, millest kõnelesime esimeses peatükis, niisiis milles seisneb kompuuterlingvistika peamine praktiline väärtus.

Lisaks sellele püüdsime näidata ka kompuuterlingvistika enda taotluste loogilist arengut lihtsate infootsisüsteemide loomise juurest keelt võimalikult sügavamalt mõistvate süsteemideni. Selle arengu käigus on arvutilingvistika muutunud puhtast rakendusteadusest üha teoreetilisemaks teaduseks. Siinses lõppsõnas tahaksime esile tõsta veel üht momenti selles arengus. Kompuuterlingvistika huvid ja meetodid põimuvad aina suuremal määral ühelt poolt teoreetilise keeleteaduse omadega, teiselt poolt aga psühholoogia omadega. Niimoodi on kompuuterlingvistika, eriti selle keelt mõistvate süsteemide loomisega tegelev haru võib-olla isegi mõneti ootamatult kujunenud keeleteaduse ja psühholoogia ühenduslüliks, mis seob neid teadusi märksa tugevamini ja orgaanilisemalt, kui see on õnnestunud nende seniste otsekontaktide kaudu. Asi on selles, et seni puudus keeleteaduse ja psühholoogia kontaktidel kindel vorm — liiga erinevad on nende teaduste uurimismeetodid. Keelt mõistvad süsteemid kui keele mõistmise protsessi mudelid on nüüd osutunud sobivaks ühisvormiks. Öeldu pole sugugi liialdus: teadlased räägivad uuest interdistsiplinaarsest uurimisalast, mis kujuneb psühholoogia, tehisintellekti ja keeleteaduse piirimail. Seda on hakatud nimetama kogitoloogiaks (lad. k. *cogitare* 'mõtlema, arutlema') ning tema ülesanne on inimese mõtlemis- ja tunnetusprotsesside, sealhulgas keele mõistmise protsesside uurimine ja nende protsesside mudelite loomine. See, missuguseks kujuneb kompuuterling-

vistika meie poolt kirjeldatud suuna, eriti aga kompuuterlingvistika kui teoreetilise uurimisala ilme tulevikus, sõltub kahtlemata suurel määral sellest, kuidas areneb eespool nimetatud teaduste sümbioos.

Kuid nii või teisiti — kompuuterlingvistika pole enam ala, kus tegeldakse üksnes tekstide automaatse töötlemise praktiliste probleemidega. Tähelepanu keskpunkti on tõusnud kõnelemise ja kõne mõistmise protsessid ning neid modelleerivate küberneetiliste süsteemide loomine. Et kompuuterlingvistika areng just selliselt on kulgenud, on igati seaduspärane. Ükskõik mis aspektist keelt uurima ei hakata, ikka jõutakse lõpuks keele välise vormi, kõne või tekstide juurest nende mõistmise sisuliste protsesside juurde. Siin on peidus keele saladused ja huvi nende vastu näib olevat inimliku uudishimu üks kõige sügavamalt juurdunud vorme. Selle kinnitusena tundub olevat sobiv tuua raamatu lõpetuseks küberneetika isa N. Wieneri mõtteavaldus: «Kokkuvõttes — inimese huvi keele vastu näib olevat kaasasündinud huvi kodeerimise ja dekodeerimise vastu, ja see näib olevat inimesele nii omane, kui üks huvi üldse saab olla. Kõne on inimese suurim huvi ja kõige eriomasem saavutus.»[47]

---

[47] Wiener N. Küberneetika ja ühiskond. Tln., 1969, lk. 75.

Kirjandust

1. peatükk. Populaarse ülevaate infokriisi erinevatest tahkudest saab raamatust *Vladimirov S., Karev M. Informatsioon ja meie. Tln., 1972.*

Teoreetilisema sissejuhatusena kompuuterlingvistika probleemidesse võib nimetada raamatut *Пиотровский Р. Г. Инженерная лингвистика и теория языка. Л., 1979.*

TRÜ keelestatistika rühma tööd on avaldatud peamiselt TRÜ kogumike sarjas «Linguistica», mida 1982. aastaks oli ilmunud juba 14 numbrit.

TRÜ tehisintellekti probleemgrupp annab välja kogumike sarja üldpealkirja all *«Труды по искусственному интеллекту»*. 1982. aastaks oli ilmunud 5 numbrit.

2. peatükk. Süvendavaks lugemiseks informaatika ja eriti infootsisüsteemide loomise probleemide kohta sobib teos *Informaatika ABC / Toim. U. Agur. Tln., 1980.*

3. peatükk. Ulatuslikumaid populaarseid ülevaateid masintõlkest pole eesti ega vene keeles viimasel ajal ilmunud. Tõsisemate huvidega lugejale sobib aga hästi raamat *Кулагина О. С. Исследования по машинному переводу. М., 1980.* Siit leiab ka ülevaate masintõlke praegusest olukorrast ja selle pakilisematest probleemidest maailmas.

Huvi — ehkki võib-olla ainult teadusajaloolist — võiksid pakkuda ka Eesti uurijate tööd masintõlke algusaegadest: *Kaasik Ü., Korjus A. Automaatsest tõlkimisest. — Keel ja Kirjandus 1959, nr. 11, lk. 663—673; Kaasik Ü., Korjus A. Automaatse tõlkimise keeleteaduslikke küsimusi. — Keel ja Kirjandus, 1960, nr. 9, lk. 545—555; Kaasik Ü., Akkel T., Korjus A. jt. Matemaatilise teksti automaatsest tõlkimisest vene keelest eesti keelde. — Kogumikus: Loodus ja matemaatika, nr. 2. Tln., 1960, lk. 102—112.*

4. peatükk. Tehisintellekti, sealhulgas ka keelt mõistvate süsteemide kohta on eestikeelset kirjandust vägagi napilt. See-eest on vene keeles viimastel aastatel ilmunud ridamisi nii populaarseid kui ka tõsiteaduslikke raamatuid. Populaarse ülevaate kõnealusest alast saab näiteks raamatuist *Рафаэл Б. Думающий компьютер. М., 1979* ning *Вейценбаум Дж. Возможности вычислительных машин и человеческий разум. От суждений к вычислениям. М., 1982.*

Keelt mõistvate süsteemide loomisega seotud probleemidesse süvenemiseks võib eelkõige soovitada kahe tuntud teose tõlkeid vene keelde: *Виноград Т. Программа понимающая естественный язык. М., 1976; Шанк Р. Обработка концептуальной информации. М., 1980.* Lisaks sellele järgmine äsja ilmunud teos: *Попов Э. В. Общение с ЭВМ на естественном языке. М., 1982.*

Aine- ja nimeregister

ADS (masintõlkesüsteem) 90
aktandiseos (sõnade vahel) 79—85
algoritmvara (masintõlkesüsteemil) 70
ammendavus (infootsingul) 43—44
AMPAR (АМПАР, masintõlkesüsteem) 89
ANASIN (АНАСИН, automaatrefereerimissüsteem) 52
andmemälu 126
andmeotsisüsteemid vt. faktiotsisüsteemid
antonüümid 37
Apresjan, J. (nõuk. lingvist) 81, 89
Armeenia NSV Teaduste Akadeemia Arvutuskeskus 67
arvutilingvistika vt. kompuuterlingvistika
assotsiatiivseosed (mõistete vahel) 36, 38—39, 43, 46, 54
automaadid 15
automaatannoteerimine 27
automaatindekseerimine 34, 41
automaatne morfoloogiline analüüs 42, 71
automaatrefereerimine 27, 51—53
automaatsõnastik 29
automaattõlge vt. masintõlge
automatiseerimine 14—15
automatiseeritud infootsing 5
— infosüsteem 32
— infotöötlussüsteem 27

Bar-Hillel, Y. (USA lingvist) 62—63
BASEBALL (keelt mõistev süsteem) 98—99, 101
Bobrow, D. (USA küberneetik) 103
Booth, A. D. (USA arvutiteadlane) 58

California Ülikool 130
Colby, J. (USA küberneetik) 103
CYRUS (keelt mõistev süsteem) 131

deskriptor 35, 37, 39—43
deskriptorkeel 33, 35, 54—55
Dewey, M. (USA bibliograafiateadlane) 33
dialoog (inimese ja arvuti vahel) 17, 19, 91, 101
dialoogotsing 49
dialoogsüsteemid 28—29, 49
dokumendiotsisüsteem 53—54
dokument 32
Dostert, B. (USA masintõlkespetsialist) 66
Drucker, P. (USA juhtimisteadlane) 8

Eesti Informatsiooni Instituut 46, 48
Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut 29
Eesti Vabariiklik Automatiseeritud Teadus- ja Tehnikainformatsiooni

Süsteem (Eesti VATTIS) 46, 48—49
ekvivalentsusseos (sõnade vahel) 36—37, 46
elemendi-hulgaseos (mõistete vahel) 113—114
elementaarteod 110, 120
elementaartähendused 106—116, 119
ELIZA (keelt mõistev süsteem) 91—92, 103, 128
elliptilisus (keeles) 23, 25, 33, 118
episoodiline mälu 126
ETAP (ЭТАП, masintõlkesüsteem) 89
Euronet DIANE (infosüsteem) 50—51

faktiotsisüsteemid 39, 53—57, 92
faktograafilised infootsisüsteemid vt. faktiotsisüsteemid
formaalkeeled 21
fraasistruktuuresitus (süntaktilisel struktuuril) 76—77
fraseoloogiline väljend 24

Georgia Tehnoloogiainstituut 131
grammatiline analüüs (lausel) 99
Green, B. (USA küberneetik) 98
Grenoble'i Ülikool 89—90
Gruusia NSV Teaduste Akadeemia Juhtimissüsteemide Instituut 67

haiguste diagnoosimine 96
Halliday, M. A. K. (inglise lingvist) 128
Heidelbergi Ülikool 89
hierarhilised seosed (mõistete vahel) 36, 40, 113
homonüümia 33, 60
Hongkongi Ülikool 89

ideede tootmine 8
ideograafiline sõnastik 35
indekseerimine 34—35, 41
infokeeled 33—34, 40, 46, 54
infokeskus 45—46
infokriis vt. infoplahvatus
infootsing 30
— ammendavus 43—44
— täpsus 43—44
infootsisüsteemid 27, 29—34, 41—44, 65, 92, 98
infoplahvatus 8—10, 14—19, 44
informaatika 27, 31—32
informatsiooni hind 10
— inflatsioon 9
— tootmine 8, 10
«Informelektro» (nõuk. infokeskus) 77
infosüsteem 31
infoteenindussüsteem 31—32
infotöötlusprotsessid 15, 19
intellektuaalne infosüsteem 105
— infotöötlus 15, 28
intelligentne süsteem 95
ISDS (rahvusvaheline perioodikaväljaannete infosüsteem) 50
ISSN (rahvusvaheline perioodikaväljaande registreerimisnumber) 50

Jeršov, A. (nõuk. matemaatik) 18
JURIOS (infootsisüsteem) 29
järeldusreeglid 119—121
järelduste tuletamine 99—101, 118, 128

Kaasik, Ü. (eesti matemaatik) 29
Karasjov, S. (nõuk. informaatik) 52
keele automaatanalüüs 22, 57
— automaatsüntees 57
— automaattöötlus 20—27, 51, 58
— morfoloogiline analüüs 56, 60, 64—65, 68, 70—75, 115
— semantiline analüüs 56, 63, 65, 68—70, 98—101, 115—116
— süntaktiline analüüs 56, 64—65, 68, 75—87, 98, 115
keeleline kommunikatsioon 20
keelevara (masintõlkesüsteemil) 50—51, 70, 86
keelt mõistvad süsteemid 28, 35, 65, 91—131
Kioto Ülikool 89
Kiušiu Ülikool 89
KLAUS (keelt mõistev süsteem) 131
kogitoloogia 132
Kolodner, J. (USA küberneetik) 130
kompuuterlingvistika 5, 27
kontekst 26, 63, 122
kujundite visuaalne äratundmine 96—97
kujundlikud väljendid 24—25
Kulagina, O. (nõuk. lingvist) 58
kvaasimõiste 37
kvaasisünonüümid 36—37
kümnendliigitus 33—34

lause grammatiline analüüs 99
— morfoloogiline analüüs 56, 60, 64—65, 68, 70—75, 115
— — esitus 69, 74—75
— semantiline analüüs 56, 63, 65, 68—70, 98—101, 115—116
— — esitus 69—71, 106—117
— — struktuur 22, 65, 106—107, 120
— süntaktiline analüüs 56, 64—65, 68, 75—87, 98, 115
— — esitus 69, 71, 75, 77, 98, 106—107, 115—116
— — struktuur 63, 65, 79, 83, 85, 101, 106—107
Lem, S. (poola kirjanik ja filosoof) 13—14, 105
Leningradi Riiklik Ülikool 67
liigitusindeksid 34—35
liigituskeeled 33
liigitussüsteemid 30
Lindsay, R. (USA küberneetik) 98—99
lingvistiline protsessor 98
Ljapunov, A. (nõuk. matemaatik) 58
LOGOS (masintõlkesüsteem) 68
loogikalis-semantilised vahekorrad (tekstis) 102
loogikasõsed 43
Luhn, H. P. (USA informaatik) 52
lähtekeel 60, 86

maailmamudel 23, 101, 103
magnetlint-andmebaasid 48—49
majandusinformatsioon 9
MARGIE (keelt mõistev süsteem) 119, 121, 128—129
masinrefereerimine vt. automaatrefereerimine
masintõlge 5, 22, 24, 27, 50, 57—90
masintõlke kvaliteet 89—90
— lähtekeel 60, 86
— väljundkeel 60, 86
masintõlkesüsteemid 28, 58—90, 92, 99, 115
— algoritmvara 70
— analüüsialgoritm 69
— keelevara 50—51, 70, 86
— morfoloogiakomponent 70—75
— põlvkonnad 65
— süntaksikomponent 70, 77—87
— sünteesialgoritm 69, 87
Massachusettsi Tehnoloogiainstituut 127
Meltšuk, I. (nõuk. lingvist) 22
metafoorid 24
Minsky, M. (USA küberneetik) 94, 123—124
mitmetähenduslikkus 21—22, 25—26, 60—63
morfoloogiakomponent (masintõlkesüsteemil) 70—75
morfoloogiline analüüs 56, 60, 64—65, 68, 70—75, 115
— esitus (lausel) 69, 74—75
— tasand 70—75
morfoloogilised kategooriad 71, 74
Muhhin, I. (nõuk. lingvist) 58
muusikat loovad süsteemid 94—95
mängivad süsteemid 94—95

Neumann, John v. (USA matemaatik) 94
NSV Liidu Teaduste Akadeemia Matemaatika Instituut 58
— Orientalistika Instituut 89
— Rakendusmatemaatika Instituut 67, 89
— Täppismehaanika ja Arvutustehnika Instituut 58, 67

Oettinger, A. (USA arvutiteadlane) 20
osa-tervikuseos (mõistete vahel) 101—102
otsieeskiri 43
otsitunnustik 41, 43

PAM (keelt mõistev süsteem) 130
Panov, D. (nõuk. matemaatik) 58
paradigma 54
paradigmaatilised seosed (mõistete vahel) 54—55, 113—114
Patendiinformatsiooni Teadusliku Uurimise Keskinstituut 67
PHRAN (keelt mõistev süsteem) 130
«Poisk» (teadustootmiskoondis) 48
polüseemia vt. mitmetähenduslikkus
Pospelov, D. (nõuk. küberneetik) 95—96
programm 18—19
programmeerimiskeeled 18—19
päring 32—33, 40—43
päringurežiim (infootsisüsteemis) 46
püsipäring 46

Quillian, M. Ross (USA küberneetik) 104

raalrefereerimine vt. automaatrefereerimine
raamatukogud 30—31, 34
Rahvusvaheline Teadus- ja Tehnikainformatsiooni Keskus 49
Raphael, B. (USA küberneetik) 94, 101
referaatajakirjad 48, 51
refereerimine 17
retro-andmepangad 49
retrospektiivne otsing 49
Riiklik Automatiseeritud Teadus- ja Tehnikainformatsiooni Süsteem 46, 49
robotid 17, 95—96, 127
Roget, M. R. (inglise filoloog) 36

Saarimaa Ülikool 89
SAD-SAM (keelt mõistev süsteem) 98—101, 118
sagedussõnastik 29
SAM (keelt mõistev süsteem) 129—130
Schank, R. C. (USA küberneetik) 110, 119—121, 124, 126, 128—130
semantikakomponent (keelt mõistval süsteemil) 106
semantiline analüüs 56, 63, 65, 68—75, 98—101, 115—116
— ellips 23
— esitus (lausel) 69—71, 106—117
— mälu 104, 126
— struktuur (lausel) 22, 65, 106—107, 120
— valents 113
— varustus (keelt mõistval süsteemil) 118
semantilised elemendid 106
— seosed 36, 39, 112
— tunnused (sõnal) 77—81
— võrgud 109, 121
SHRDLU (keelt mõistev süsteem) 127—129
SIMPAR (СИМПАР, masintõlkesüsteem) 89
SIR (keelt mõistev süsteem) 101—103
soo-liigiseos (mõistete vahel) 37—40, 43, 46, 54, 113
Spengler, O. (saksa filosoof) 13
Stanfordi Rahvusvaheline Uurimisinstituut 131
strukturaallingvistika 105
stsenaariumid 129
STUDENT (keelt mõistev süsteem) 103
suhtlemine (inimese ja arvuti vahel) 92
sõltuvusesitus (süntaktilisel struktuuril) 76—87, 128
sõltuvusmall (sõnal) 77—81
sõnaühendid 22, 26, 78
sünonüümia 22, 25, 33, 36
süntagmaatilised seosed (keeleüksuste vahel) 54, 113
süntaksikomponent (masintõlkesüsteemis) 70, 77—87
süntaksireeglid 60
süntaktiline analüüs 56, 64—65, 68, 75—87, 98, 115
— esitus (lausel) 69, 71, 75, 77, 99, 106—107, 115—116
— struktuur (lausel) 63, 65, 79, 83, 85, 101, 106—107
— valents 116
süntaktilised seosed 81—87
— tunnused (sõnal) 77—87
süstemaatiline kataloog 30—31
SYSTRAN (masintõlkesüsteem) 64, 88—89

Žolkovski, A. (nõuk. lingvist) 22

Tartu Riikliku Ülikooli keelestatistika uurimisrühm 29
— tehisintellekti probleemgrupp 29
Taube, M. (USA informaatik) 59
teadmisskeemid 123—126
teadus- ja tehnikainformatsiooni keskus 45
teadus- ja tehnikarevolutsioon 11, 14
teaduse areng 11—12
teaduse ja tehnika progress 11—13
TEAM (keelt mõistev süsteem) 131
tehisintellekt 28, 93—132
tehisintellektisüsteemid 65
teksti modelleerimine 5
tekstinformatsioon 16—17
tekstitöötlus 32
tesaurus 29, 35—43, 50, 55
TITUS (masintõlkesüsteem) 50
Toffler, A. (USA sotsioloog) 10
Turing, A. M. (inglise matemaatik) 91
tõenäosuslikud järeldused 121
Tõugu, E. (eesti arvutiteadlane) 20
tähendus 21, 100, 105—122
täpsus (infootsingul) 43—44

UDK (universaalne kümnendliigitus) 30—31, 33—34
UNESCO 49
UNISIST (rahvusvaheline teadus- ja tehnikainfo koostööprogramm) 49

valents (sõnal) 78—79
valikteadistus 46
varieeruvus (keelel) 37
väljundkeel 60, 86

Weaver, W. (USA arvutiteadlane) 58
Weizenbaum, J. (USA küberneetik) 91—92, 103
Wiener, N. (USA küberneetik) 94, 133
Wilensky, R. (USA küberneetik) 130
Wilks, Y. (USA küberneetik) 110—112
Winograd, T. (USA küberneetik) 127—129

Üleliiduline Teadus- ja Tehnikainformatsiooni Instituut 48, 51
Üleliiduline Teadus- ja Tehnikakirjanduse ning Dokumentatsiooni Tõlkekeskus 89

Yale'i Ülikool 130

Sisukord

Saateks 5

1. INIMENE JA INFORMATSIOON 7

1.1. Inimene ja informatsioon — poolfilosoofiline sissejuhatus 7
1.2. Infokriis, arvutid ja keel — keeleline sissejuhatus 16

2. TEKSTIDE OTSING JA TÖÖTLUS AUTOMATISEERITUD INFOSÜSTEEMIDES 30

2.1. Automatiseeritud infootsisüsteemid — probleemid ja ehitus 30
2.2. Automatiseeritud infootsisüsteemide arengutendentsid 51

3. MASINTÕLGE 57

3.1. Ajalugu ja probleemid 58
3.2. Kuidas töötab masintõlkesüsteem 68
3.3. Mõni sõna masintõlkest tänapäeva maailmas 88

4. KEELT MÕISTVAD SÜSTEEMID 91

4.1. Keele mõistmine ja tehisintellekt 93
4.2. Esimesed süsteemid ja põhimõttelised probleemid 98
4.3. Semantiline analüüs ja teadmiste semantilised mudelid 104
4.4. Keelt mõistvaid süsteeme 1970-ndatest aastatest 127

Lõpetuseks 131
Kirjandust 133
Aine- ja nimeregister 135

Серия «Мозаика».
Халдур Ыйм.
Человек, язык и ЭВМ или компьютерная лингвистика.
На эстонском языке.
Под редакцией У. Агур.
Художник-оформитель Т. Ару.
Таллин, «Валгус».
Toimetaja H. Seppo.
Kunstiline toimetaja R. Eilsen.
Tehniline toimetaja E. Sagris.
Korrektorid E. Vernik, H. Luik.
ИБ № 1447.

Laduda antud 04. 03. 81.
Trükkida antud 01. 03. 83.
MB-00531.
Formaat 60×84/16.
Trükipaber nr. 1.
Kiri Literaturnaja.
Kõrgtrükk.
Tingtrükipoognaid 8,37.
Tingvärvitõmmiseid 8,73.
Arvestuspoognaid 9,54.
Trükiarv 10 000.
Tellimuse nr. 724.

Hind 50 kop.

Kirjastus «Valgus», 200090 Tallinn, Pärnu mnt. 10.
Trükikoda «Ühiselu», 200001 Tallinn, Pikk 40/42.

SARJAS «MOSAIIK»
SENI ILMUNUD:

Nr. 1. Valeri Soifer
PÄRILIKKUSE
ARITMEETIKA

Nr. 2. Robert Wells
BIOONIKA

Nr. 3. Samuel Lilley
INIMESED, MASINAD
JA AJALUGU

Nr. 4. Leonard Rastrigin
JUHUSLIK, JUHUSLIK,
JUHUSLIK MAAILM

Nr. 5. George F. Bass
ALLVEEARHEO-
LOOGIA

Nr. 6. Vladimir Levi
MÕTTEJAHT

Nr. 7. Haldur Õim
SEMANTIKA

Nr. 8. Boriss Sergejev
HUVITAVAT
FÜSIOLOOGIAST

Nr. 9. Ago Künnap,
Paula Palmeos,
Tõnu Seilenthal
PÕHJA JA ITTA

Nr. 10. Samuel Tolansky
REVOLUTSIOON
OPTIKAS

Nr. 11. Aron Gurevitš
VIIKINGITE RETKED

Nr. 12. James E. Gordon
MIKS ME EI KUKU
LÄBI PÕRANDA

Nr. 13. Mihhail
Volkenštein
TEADUSE RISTTEEDEL

Nr. 14. Alan Holden
MIS ON TAHKE KEHA

Nr. 15. Viktor Komarov
UNIVERSUMI
SÜGAVUSTES

Nr. 16. Nikolai
Tšeboksarov,
Irina Tšeboksarova
RAHVAD, RASSID,
KULTUURID

Nr. 17. Marell Minnaert
VALGUS JA VÄRV
LOODUSES

Nr. 18. Boriss Sergejev
IMETABANE AJU

Nr. 19. Pierre-André
Moléne
TAIFUUNIKÜTID

Nr. 20. David Finkelštein
AINE PUHTUS

Nr. 21. Viktor Pekelis
SINU VÕIMALUSED,
INIMENE

Nr. 22. Jakov Hurgin
AGA MIS SIIS?

Nr. 23. Jevgeni Gruntenko
IMMUUNSUS—
ORGANISMI LIITLANE
VÕI VAENLANE?

Nr. 24. Peter Fischer,
Reinhard Göttner,
Reinhard Krieg
MIS ON JA MIDA
SUUDAB STATISTIKA

Nr. 25. Bernard Glemser
INIMENE
VÄHI VASTU

Nr. 26. Friedrich Zavelski
MASS JA SELLE
MÕÕTMINE

Nr. 27. Kalojan Manolov,
Dobri Lazarov, Ivan Lilov
KEEMIAL ON OMA
SEADUSED

Nr. 28. Rafail Guter, Juri Polunov
ABAKUSEST RAALINI

Nr. 29. Ritchie R. Ward
BIOLOOGILINE KELL

Nr. 30. Roland Glaser
TEISITINÄHTUD BIOLOOGIA

Nr. 31. Aino Lunge
EMOTSIOONIDE PSÜHHOLOOGIA

Nr. 32. Jossif Šklovski
UNIVERSUM, ELU, MÕISTUS

Nr. 34. Stephen Barr
KÄÄRIDE JA PABERIGA TOPOLOOGIASSE

LÄHEMAL AJAL ILMUVAD:

Nr. 35. Henry Lipson
KATSED, MIS LÕID FÜÜSIKA

Nr. 36. Vitali Rõdnik
AATOMIMAAILMA SEADUSED

286.413

50 kop.

33 mosaik

Keel on inimese evolutsiooni suursaavutus, imepärane tööriist maailma tunnetamiseks ja suhtlemiseks inimeste ühistegevuses.

Keele varjatud mehhanismide keerukust hakkame paremini mõistma, püüdes õpetada keelest aru saama inimvaimu loomingu enda suursaavutust — arvutit. Keele automaatanalüüsi ja sünteesi probleemidest ongi juttu käesolevas raamatus.

EESTI RAHVUSRAAMATUKOGU
Z-101910