Keele automaattöötlus ja automatiseeritud infosüsteemid

HALDUR ÕIM

MÕNI SÕNA TERMINOLOOGIAST. Keele automaattöötluse all mõeldakse siin keelematerjali, eelkõige tekstide töötlemist elektronarvuti abil mitmesugustel eesmärkidel. Selle tegevus- ja uurimisala nimetamiseks pole eesti keeles kindlat terminit juurdunud juba sellepärast, et sel teemal pole eesti keeles eriti palju kirjutatud. Termin *rakenduslingvistika* on liiga lai, sest mitte igasugune keeleteaduse tulemuste rakendamine või igasugune keelematerjali töötlemine rakenduslikel eesmärkidel ei toimu elektronarvutite abil. Inglise keeles tehakse näiteks vahet terminite *computational linguistics* ja *applied linguistics* vahel. Ainult esimene neist seostub meid siin huvitava uurimisalaga, teine tähistab aga mitmesuguseid muid rakenduslikke keeleuuringuid, eriti näiteks keelte õpetamisega seotud uuringuid. Vene keeles kasutatakse inglise oskussõna *computational linguistics* ligilähedase vastena tervet rida termineid: *прикладная лингвистика, инженерная лингвистика, вычислительная лингвистика, машинная лингвистика.*

Enne kui hakata kaaluma, missugune eestikeelne termin sobiks nende vasteks, tuleb osutada veel ühele aspektile nende tarvituses. Kui me kasutame väljendit *keele automaattöötlus*, siis mõtleme selle all keelematerjali automaatset töötlemist kui tegevusala. Ülaltoodud terminid haaravad aga eelkõige teoreetilist laadi teaduslikke uurimusi, mis moodustavad erilise lingvistikaharu (või vähemalt hübriidteaduse, mille üks komponent pärineb lingvistikast). See vahetegemine pole küll väga kindel, eeltoodud termineid kasutatakse sageli vahet tegemata nii teoreetilistest uurimustest kui ka praktilistest keele automaattöötlussüsteemidest kõneldes; kuid teiselt poolt on nii inglise kui ka vene keeles olemas terminid, mida kasutatakse, kui tahetakse osutada just praktilisele tegevusalale: vastavalt *automated language processing* ja *автоматическая обработка текстов.* Just nende vasteks on käesolevas artiklis mõeldud termin *keele automaattöötlus.*

Missuguse terminiga peaksime tähistama vastavat «poollingvistilist» teaduslikku distsipliini, mis on keele automaattöötluse kui tegevusala aluseks, varustades teda sobivate keelekirjeldusmeetodite, üksikuurimuste ja teoreetiliste kontseptsioonidega? Keele automaattöötlus on laiemast aspektist vaadates üks liik automaatset infotöötlust. Keelelisi tekste töödeldes on enamasti eesmärgiks töödelda informatsiooni, mille kandjaks need tekstid on. Seepärast oleks meie arvates sobiv vastavat lingvistikaharu nimetada i n f o t ö ö t l u s l i n g v i s t i k a k s või lühemalt — ja ka kõlavamalt — i n f o l i n g v i s t i k a k s. Eeskujud psühho-, sotsio-, neuro- jne. lingvistika näol on olemas.

Eeltoodud vahetegemise alusel võime nüüd öelda, et käesolevas artiklis tuleb põhiliselt juttu keele automaattöötlusest. Räägime sellest, milles töötlemine seisneb, kuidas ja mis eesmärkidel töödeldakse. Siinseal peatume ka teoreetilisematel infolingvistika probleemidel.

KEELE AUTOMAATTÖÖTLUSE AJALOOST JA TEMA TAUSTAST TÄNAPÄEVAL. Keele automaattöötlus sündis teatavasti masintõlkimisena 1950-ndate aastate algul. Millegipärast tundus just tekstide tõlkimine elektronarvuti abil eriti ahvatlev ja paljutõotav. Lühikese ajaga sai masintõlkimisest üks populaarsemaid küberneetilisi probleeme. Sellega tegelevaid töörühmi ja laboratooriume tekkis kõikjal, ka Nõukogude Liidus (sealhulgas meil Eestiski). Esialgne buum möödus siiski võrdlemisi kiiresti. Jõuti veendumusele, et masintõlke probleemi pole võimalik lahendada keele enda ehitust põhjalikult tundma õppimata. Keel osutus lähemal uurimisel aga äärmiselt keerukaks ja mitmetahuliseks nähtuseks. Nii pidi ka USA-s asja uurinud riiklik komisjon oma aruandes konstateerima, et töö jätkamisel senises stiilis ei ole perspektiivi ning et probleemi edukaks lahendamiseks tulevikus on tarvis põhjalikke uurimusi keele ehituse ja funktsioneerimise kohta.[1]

Öeldu ei tähenda siiski, nagu oleks masintõlkimise idee täielikult maha maetud. Selle probleemiga tegeldakse endiselt paljudes keskustes üle maailma, ehkki keegi ei looda selle lõplikku lahendamist lähemate aastate jooksul.[2]

Kuid masintõlkimine ei ole keele automaatses töötlemises praegu enam ainus ja kaugeltki mitte ka peamine ala. Ameerika tuntumaid infolingvistika spetsialiste D. G. Hays on öelnud: kui keele automaattöötluses tegeldaks ainult masintõlkega, oleks meil tegu kõrbega, kus on küll rohkesti miraaže, kuid vähe oaase.[3] Tegelikult käib kõnealusel alal vilgas elu, mis on suhteliselt sõltumatu masintõlkest.

Tõepoolest, tekstide tõlkimine ühest keelest teise on vaid üks ja seejuures kindlasti mitte olulisim probleem suurest probleemikompleksist, mis on seotud kõikvõimalike tekstide hulga kiire kasvuga. Neid tekste on vaja nii või teisiti töödelda. On hulk probleeme, mis palju pakilisemalt ootavad lahendust ja mille lahendamisel elektronarvutid võivad (praegu) anda palju tõhusamat abi kui tõlkimisel. Need on mitmesugused probleemid, mille puhul kasutatakse üldist mõistet *infotöötlus*. Just selle mõiste alla mahub praegu valdav osa keele automaattöötluse alal tehtavast tööst. Siit tulenevad peamised praktilised tellimused ja see inspireerib ka suuremat osa teoreetilisi uurimusi. Selles kontekstis vaatleme siinkohal keele automaattöötlust ka meie.

Niimoodi mõistetud automaatses keelematerjali töötluses võib esile tuua kaks probleemiringi. Esiteks: nõutava informatsiooni *otsimine* algselt keeleliste tekstidena vormistatud ja raali mällu viidud massiivist. Selleks otstarbeks loodavaid süsteeme nimetatakse automatiseeritud infootsisüsteemideks (lühendatult IOS). Teiseks: keelelise informatsiooni automaatne *töötlemine* mitmesugustel eesmärkidel (s. o. mitte lihtsalt ülesotsimine), eriti informatsiooni sisuline, semantiline töötlemine.

Nende kahe suuna keskset osa on rõhutanud paljud autorid. Näiteks

---

[1] Language and Machines, Computers in Translation and Linguistics. Publication 1416. Automatic Language Processing Advisory Committee Report. Washington, D. C., 1966. Selle aruande venekeelne tõlge koos kommentaaridega leidub: Научно-техническая информация. 1968, сер. 2, nr. 8, lk. 23—36.

[2] Mõninga ülevaate masintõlke alal saavutatust annab kogumik: Автоматический перевод. Москва, 1971; eriti informatiivne on selle ulatuslik sissejuhatus I. Meltšukilt ja O. Kulaginalt: Автоматический перевод: краткая история, современное состояние, возможные перспективы (lk. 3—25).

[3] D. G. Hays, Applied Computational Linguistics. Applications of Linguistics. Selected Papers of the Second Congress of Applied Linguistics. Cambridge, 1969. Ed. by G. E. Perren and J. L. M. Trim. Cambridge, 1971, lk. 69—85.

H. Borko, üks tuntumaid keele automaattöötluse spetsialiste USA-s, nimetab oma ülevaates järgmisi põhilisi suundi: a) tõlkimine ühest keelest teise; b) informatsiooni talletamise ning otsimise meetodite edasiarendamine; c) «intelligentsete» automaatide loomine, mis oleksid suutelised vastama küsimustele loomulikus keeles.[4]

Masintõlkimisest ja selle kohast keele automaattöötluses tänapäeval me juba rääkisime pisut. Ülejäänud kaks suunda aga ühtivad sisuliselt nendega, mida mainisime ülemal. Ainult kui H. Borko punktis c räägib automaatidest, mis vastavad küsimustele, on see liiga kitsas iseloomustus. Keelt töötlevad süsteemid, mille loomisega sel alal tegeldakse, ei ole mõeldud ainult küsimustele vastamiseks, vaid ka mitmeks muuks otstarbeks: loogiliste järelduste tuletamiseks lausetest, loomulikus keeles antud käskude täitmiseks jm. Seepärast on täpsem rääkida siin keelelist informatsiooni semantiliselt töötlevatest süsteemidest e. semantilistest infotöötlussüsteemidest.

Alljärgnevas tutvustamegi, mida kahel mainitud alal on korda saadetud.

Automatiseeritud infootsisüsteemid

«Infoplahvatusest» ja «infokriisist» on viimasel ajal palju räägitud. Maailmas ilmub iga aastaga aina rohkem trükiseid: teaduslikke ja populaarteaduslikke artikleid ning raamatuid, aparaatide, tehnoloogiliste protsesside, leiutiste, avastuste jm. kirjeldusi, millest vastava ala spetsialistil peab olema ülevaade. Andmeid selliste trükiste hulga kasvu kohta on küllalt sageli esitatud, mistõttu pole mõtet neid siin korrata. Võib tuua ainult ühe kujuka näite selle kohta, mis infouputus maksma läheb. H. Borko väidab (H. Humprey andmeid aluseks võttes), et USA-s on teaduslike uurimistööde asjatule dubleerimisele, mis on olnud tingitud puudulikust infovahetusest, kulunud 300 000 inimaastat, kusjuures neile uurimistöödele on subsiidiumide näol raisatud 2 biljonit dollarit.[5]

Mehhaniseerimise ja automatiseerimise abil on inimestel õnnestunud muuta ühiskonna praktilise, majandusliku tegevuse areng sõltumatuks oma otsestest füüsilistest võimetest. Selle tagajärjel aga on ühiskonnaelu see külg paisunud niivõrd ulatuslikuks, et inimesed ei tule enam toime ka informatsiooniga, mis kajastab siin kulgevaid protsesse ja millest lähtudes nad seni on neid protsesse valitsenud. Väljapääs on ainult üks: on vaja automatiseerida ka informatsiooni töötlemine.

Automatiseeritud infootsisüsteemid ongi mõeldud lahendama üht osa sellest probleemist. Võib ka kohe öelda, et see on seni ainus tüüp keele automaattöötluse süsteeme, mis on jõudnud eksperimentaalstaadiumist välja ning on hakanud juurduma praktikas.

Automatiseeritud IOS luuakse suurte tekstimassiivide talletamiseks ja vajaminevate tekstide kiireks ülesotsimiseks sellisest massiivist. IOS-i baasil rajatakse enamasti terviklik infokeskus, mis võib täita mitmesuguseid ülesandeid. Kirjanduses räägitakse näiteks kolme tüüpi infokeskustest: 1) automatiseeritud raamatukogu või -hoidla tüüpi infokeskused, mis hangivad, selekteerivad ja talletavad teatavat liiki tekste ning võimaldavad klientidel neile operatiivselt juurde pääseda; 2) keskused, mis peale eelneva tegelevad referaatide, registrite, temaa-

[4] H. Borko, Automated Language Processing. New York — London, 1967.
[5] H. Borko, Automated Language Processing, lk. 9.

tiliste bibliograafianimestike ja muu sellise automaatse koostamise ning levitamisega, seda enamasti kitsama ainevalla ulatuses; 3) keskused, mis lisaks eelnevale tegelevad veel informatsiooni ja infoprotsesside automaatse analüüsimisega vastaval alal: dokumentide struktuuri ja dokumentidevaheliste seoste analüüsiga, informatsiooni liikumise kindlakstegemisega, vastava teadus- või tehnikaala arengutendentside väljaselgitamise ja prognoosimisega jne.[6]

INFOKEELED. Automaatse IOS-i töötamise aluspõhimõte on lihtne. Süsteem peab võrdlema küsimuse e. päringu teksti arvuti mälus leiduvate tekstidega ning väljastama tekstid (või andmed tekstide kohta), mis sisaldavad informatsiooni selle küsimuse alalt. Lihtsaim võrdlemise variant seisneb selles, et arvuti kontrollib, missugustes tekstides sisalduvad päringus kasutatud sõnad. Ent selline triviaalne võrdlus on arusaadavalt väga ebaefektiivne. Otseselt päringus leiduvate sõnade esinemine tekstides ei ole kaugeltki piisav tingimus, mis tagaks kõigi päringule sisuliselt vastavate tekstide äratundmise. IOS-i efektiivseks funktsioneerimiseks on tarvis süsteemi mällu viidavad tekstid tõlkida spetsiaalsesse infokeelde. Iga loodava IOS-i jaoks töötatakse välja oma infokeel, mis saadakse loomuliku keele, täpsemini vastava ala allkeele erilise töötlemise teel. Infokeele põhiliseks ülesandeks on 1) neutraliseerida loomuliku keele varieeruvus, tagada, et igale mõistele vastaks ainult üks väljend ja igale väljendile vaid üks mõiste, ning 2) fikseerida keelde kuuluvate mõistete vahel loogiliste vahekordade süsteem.

Tüübilt on infokeeli mitmesuguseid.[7] Automatiseeritud IOS-ides kasutatavatest infokeeltest on kõige levinumad nn. deskriptorkeeled, mis on tüüpiliselt vormistatud tesaurustena. Deskriptorid kujutavad endast infokeele leksikaalseid üksusi, mis tähistavad vastava ala kindlaid mõisteid ja mille kaudu fikseeritakse infokeele loogiline struktuur.

Et tesaurus on nüüdisaegsete automatiseeritud IOS-ide keskseid komponente ja et just tesauruste loomisega on seotud põhilised keelelise analüüsi probleemid IOS-ides, siis peatume neil veidi lähemalt.[8]

Tesaurus kujutab endast deskriptorite korrastatud loendit, kus iga deskriptori juures on näidatud tema loogilised seosed teiste deskriptoritega. Deskriptor, nagu öeldud, esindab põhimõtteliselt üht kindlat mõistet vastavalt alalt. Kitsamate erialade jaoks loodavates tesaurustes on seepärast deskriptoriteks eelkõige selle ala terminid. Üldjuhul pole aga mingi mõiste valimisel deskriptoriks otsustav, kas teda tähistab kindel termin või mitte. Põhiline kriteerium on see, kas mõistega seoses võidakse otsida informatsiooni või mitte, s. o. kas selle kohta

---

[6] Г. Сэлтон, Автоматическая обработка, хранение и поиск информации. Москва, 1973, lk. 17.

[7] Põhjaliku ülevaate infokeelte tüüpidest ja üksikute tüüpide konkreetsetest iseärasustest, samuti infokeelte loomise meetoditest annab: В. А. Москович, Информационные языки. Москва, 1971.

[8] Tesauruste kohta olemasolevast äärmiselt ulatuslikust kirjandusest toome siin ainult mõned olulisemad tööd, mis on otsesemalt seotud tesauruste lingvistilise aspektiga: М. В. Арапов, Некоторые принципы построения словаря типа «тезаурус». Научно-техническая информация. 1964, nr. 4, lk. 7—14; А. И. Черный, Общая методика построения тезаурусов. Научно-техническая информация. 1968, сер. 2, nr. 5, lk. 9—32; D. Soergel, Eine Einleitung zur Herstellung von Klassifikationssystemen und Thesauri im Bereich der Dokumentation. Frankfurt a. M., 1969; Д. Варга, Методика подготовки информационных тезаурусов. Сборник переводов по вопросам информационной теории и практики. ВИНИТИ. Москва, 1970; Ю. А. Шрейдер, Тезаурусы в информатике и теоретической семантике. Научно-техническая информация. 1971, сер. 2, nr. 3, lk. 21—24.

või selle abil võidakse midagi küsida. Eriti laiema temaatikaga IOS-ide puhul tuleb seepärast tesaurusesse võtta võrdlemisi laia diapasooniga sõnavara.

Tesauruse üks põhilisi ülesandeid on loomuliku keele suure varieeruvuse neutraliseerimine. Selle all mõeldakse eelkõige sünonüümia ja homonüümia välistamist.

Üht ja sama mõistet võidakse loomulikus keeles väljendada väga erisuguste sõnade ja sõnaühenditega. Seetõttu tuleb iga deskriptori juures esitada kõik vaatlusaluse mõiste tähistusviisid, mida vastava ala allkeeles kasutatakse. Näiteks võime seadusandlikes tekstides[9] leida väljendeid *joobnud, purjus, joobeseisundis, alkoholijoobes, ebakaines olekus*, mis kõik tähistavad üht ja sedasama seisundit. Kui soovime kätte saada tekste, milles on kõnealusest seisundist juttu, tuleb otsida kõiki dokumente, kus ükskõik missugune neist sünonüümidest ette tuleb.

Niisamuti on tarvis tekstides leida ja spetsiaalselt ära märkida homonüümid ning mitmetähenduslikud sõnad, mida loomulikus keeles, nagu ka mis tahes selle allkeeles, on rohkesti (ja mille kohta siin vaevalt on vaja eriti palju näiteid tuua): *kulu*$_1$ (= kuluhein) ja *kulu*$_2$ (= kulutus); *maa*$_1$ (= maatükk), *maa*$_2$ (= maismaa), *maa*$_3$ (vastandatult linnale) jne. Arusaadavalt on tarvis, et nõutud tekstide otsimisel suudaks süsteem selliseid tähendusi eristada (näiteks päringu peale, mis puudutab maa jagamist töötajatele, ei väljastaks tekste, milles räägitakse maa ja linna erinevuste kaotamisest).

Ent ka sünonüümia ja homonüümia välistamine üksi ei taga sugugi vajalikku täielikkust ja täpsust tekstide otsimisel. Seepärast on IOS-ide tesaurustes lisaks nimetatud seostele fikseeritud ka mitmesugused teised deskriptoritevahelised loogilis-semantilised seosed, mille abil saab otsimisprotseduuri täiendada ja täpsustada. Kõige olulisemad informatsiooni otsimise seisukohalt on leitud olevat soo-liigi-seosed. Igal deskriptoril on kindel koht soo- ja liigimõistete hierarhias või hierarhiates, mis tesauruses on esile toodud. Näiteks on karistus (juriidilises mõttes) üks liik mõjutamisvahendeid, kriminaalkaristus üks liik karistusi, vabadusekaotus aga üks liik kriminaalkaristusi jne. Palk on üks liik töötasusid kõrvuti selliste liikidega — ehkki erineva liigitusaluse järgi — nagu kuutöötasu, lisatöötasu, ületunnitasu jne.; töötasu omakorda on üks liik sissetulekuid kõrvuti selliste liikidega nagu pension, stipendium, preemia, honorar jt. Nende seoste tähtsus vajaliku informatsiooni otsimisel tuleneb asjaolust, et sisuliselt üht ja sedasama probleemi võib loomulikus keeles käsitleda kord üldisemate, kord konkreetsemate mõistete kaudu. Kui meid huvitab näiteks, kuidas on reguleeritud ühe või teise töötajate kategooria töötasud, siis tuleb silmas pidada, et meid huvitav materjal võib sisalduda ühelt poolt dokumentides, mis reguleerivad kõnealuste töötajate sissetulekuid üldse, teiselt poolt aga ka dokumentides, mis reguleerivad nende üht või teist töötasu liiki (palka, lisatasu, ületunnitasu jne.). Kui vastavad seosed on tesauruses (infokeeles) fikseeritud, võib IOS neid otsimisprotseduuris juba automaatselt arvesse võtta.

Lisaks soo-liigi-seostele arvestatakse enamikus tesaurustes ka nn. assotsiatiivseid seoseid. Nende seas võidakse eristada veel üksikuid kindlama sisuga seosetüüpe, näit. osa-tervik: *eluruum — elamu*; agentiivne (või laiemalt põhjuslik) seos: *lektor — loeng, naka-*

[9] Siinsed ja suur osa järgnevaid näiteid on võetud TRÜ kriminoloogia laboratooriumis loodava juriidilise IOS-i tesaurusest. Sellest süsteemist endast tuleb lähemalt juttu allpool.

tumine — haigestumine jms. Siia kuuluvad aga ka täpselt määratlemata sisulised seosed, mis põhinevad vastavate sõnadega tähistatavate esemete või nähtuste faktilistel seostel tekstide poolt kirjeldataval alal. Näiteks võime sel alusel mõistega *töötervishoid* siduda sellised mõisted nagu *ventilatsioon, tolm, vibratsioon, tervist kahjustav töö* jms.

Tesauruste ja üldse infokeelte koostamiseks on mitmeid meetodeid. Põhiline on nn. loogilis-intuitiivne meetod, mis seisneb tekstidest valitud sõnade massiivi otseses semantilises läbitöötamises inimeste poolt ning võrreldavate sõnade vahel valitsevate sisuliste seoste kindlaksmääramises. Tesauruste koostamiseks on aga pakutud ja katsetatud ka mitmeid automaatseid meetodeid.[10] Kõik need põhinevad tekstide statistilisel analüüsil. Peamine eeldus on seejuures järgmine: kui kaks sõna esinevad koos mingi intervalli — lause või suurema tekstilõigu — piires küllalt kõrge sagedusega, on nad ka sisuliselt omavahel kuidagiviisi seotud. Nende distributiiv-statistiliste meetodite põhiline puudus on selles, et nad ei võimalda hinnata leitud seoste täpset sisulist olemust. Seetõttu kasutatakse neid tavaliselt koos loogilis-intuitiivse meetodiga, näiteks algmaterjali hankimiseks viimasele või ka viimasega saadud tulemuste täiendamiseks.

INFOOTSISÜSTEEMI FUNKTSIONEERIMINE. IOS-i raames, nagu öeldud, kujutab sel viisil loodud tesaurus endast infokeelt, millesse süsteemi mällu viidavad tekstid tõlgitakse. Sellist tõlkimist nimetatakse i n d e k s e e r i m i s e k s ja see seisneb lihtsalt öeldes tekstides leiduvate sõnade ja sõnaühendite asendamises vastavate infokeele üksustega — deskriptoritega. Indekseerimine võib toimuda käsitsi, ent suurte tekstimassiivide puhul on praktiliselt mõeldav ainult automaatne indekseerimine.

Viimasega on aga seotud terve hulk keele automaatse analüüsi probleeme.

Eelkõige on vajalik tekstide a u t o m a a t n e m o r f o l o o g i l i n e a n a l ü ü s, et tekstis ära tunda mingi sõna erinevaid morfoloogilisi vorme, mis tuleb tõlkida üheks ja samaks deskriptoriks. Selleks on tarvis näit. programme käände- ja pöördelõppude eraldamiseks, samuti sama sõna eri tüvekujude äratundmiseks (eesti keeles näit. astmevahelduslike sõnade puhul).

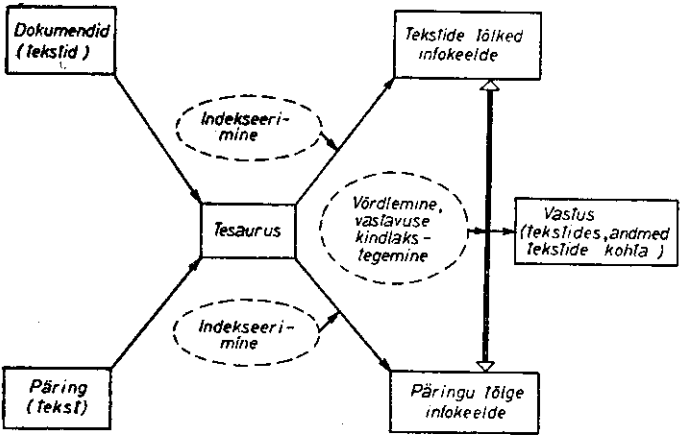
Samuti on vaja teatavas ulatuses teha a u t o m a a t s e t s ü n t a k t i l i s t a n a l ü ü s i, ja seda kahel põhjusel. Esiteks võivad paljud infokeele deskriptoriteks valitud mõisted loomulikus keeles olla väljendatud sõnaühenditega, mille automaatne äratundmine tekstis on probleem näiteks kas või sellepärast, et mingisse sõnaühendisse kuuluvad sõnad ei tarvitse tekstis alati järjestikku paikneda, teiselt poolt aga ei tähenda mingi sõnaühendi koosseisu kuuluvate üksiksõnade leidumine samas lauses veel, et meil on tegemist just selle sõnaühendiga. Eesti keeles sugeneb raskusi ka põhjusel, et mõningaid sõnu kirjutatakse kord lahku, kord kokku (näit. abimäärsõnaühendid jms.). Teiseks aitab süntaktiline analüüs homonüümide ja mitmetähenduslike sõnade puhul tähendusi eristada, nimelt neil juhtudel, kus erinevad tähendused on kontekstis olevate sõnade kindlate tunnuste põhjal äratuntavad.

---

[10] Ülevaate neist leiab: В. А. Москович, Информационные языки. Москва, 1971, IV peatükk; В. А. Москович, Дистрибутивно-статистический метод построения тезаурусов: современное состояние и перспективы. Всесоюзный семинар по информационным языкам. Предварительные публикации. Вып. I. ВИНИТИ. Москва, 1971.

Tekstide otsimine süsteemi mälust mingi kindla päringu peale toimub nüüd ainult nende infokeelde tõlgitud (indekseeritud) kujude kaudu. Päringu tekst indekseeritakse samuti, täiendades seda tesauruse põhjal vastavalt nõudmisele soo-, liigi- ja muude deskriptoritega. Saadud avaldist võrreldakse mälus leiduvate tekstide indekseeritud kujudega. Seejärel väljastatakse tekstid või andmed tekstide kohta, mille otsimiskujud nõutud viisil vastavad päringu otsimiskujule.

Ühtekokku võime IOS-i funktsioneerimise skeemi esitada järgmisel kujul.



INFOOTSISÜSTEEMIDE KASUTAMINE. Püüdsime kirjeldada automatiseeritud infootsisüsteemi kõige üldisemat ehitust, tema peamisi komponente ja töötamispõhimõtteid. Nagu eespool mainitud, tuleneb IOS-ide olulisus ja eriline koht keele automaattöötluses tänapäeval ühelt poolt asjaolust, et nad on praegu faktiliselt ainus keele automaattöötluse süsteemitüüp, mis on jõudnud praktikasse; teiselt poolt on see aga tingitud asjaolust, et need süsteemid moodustavad — lisaks nende kasutamisele otseselt informatsiooni otsimiseks — ühe keskse komponendi ka paljudes süsteemides, mis keelelisi tekste töötlevad muudel eesmärkidel (näit. tekstide automaatrefereerimisega ja üldisemalt nende sisulise kokkusurumisega tegelevad süsteemid, samuti indeksite, temaatiliste klassifikatsioonide ning nimestike jms. koostamiseks mõeldud süsteemid). Ka viimati mainitud ülesannete täitmiseks on eelkõige vaja töödeldavates tekstides identifitseerida tekstiosad (sõnad, sõnaühendid, laused), mis esindavad üht või teist teemat. Näiteks automaatrefereerimise (s. o. referaatide ja annotatsioonide automaatse koostamise) all mõeldakse põhiliselt protseduuri, kus tekstist nopitakse välja üksikud laused, nende osad või ka üksiksõnad, millel on selle teksti kui terviku seisukohalt kindel väärtus, s. t. nad esindavad selles arendatavat teemat. Relevantsete tekstiosade väljaselgitamiseks kasutatakse ühelt poolt mitmesuguseid statistilisi kriteeriume, teiselt poolt aga mõistetevahelisi seoseid, nagu need on fikseeritud tesaurustes.[11]

[11] Näit. H. P. Luhn, The Automatic Creation of Literature Abstracts. «IBM Journal of Research and Development» 1958, vol. 2, nr. 2, lk. 159—165; H. P. Edmundson, Problems of Automatic Abstracting. «Communications of ACM» 1964, vol. 7, nr. 4, lk. 13—20; И. П. Севбо, Структура связного текста и автоматизация реферирования. Москва, 1969; В. А. Аграев, Л. А. Серебрякова О возможности смысловой компрессии документов на основе извлечения предло

Veelgi ilmsem on IOS-i tüüpi süsteemi vajalikkus indeksite, aineregistrite, temaatiliste nimestike jms. automaatsel koostamisel, kus põhiliseks probleemiks on kindlaks teha teatavate mõistete esinemine tekstiosas või tekstis ja selle põhjal määrata vastava tekstiosa või kogu teksti relevantsus meid huvitava teema suhtes.

INFOOTSISÜSTEEMIDE TÄIENDAMINE JA ARENGUTENDENTSID. Eespool kirjeldatu andis ülevaate nn. klassikaliste infootsisüsteemide ehitusest ja tüüppõhimõtetest. Nende süsteemide loomise ja praktikasse juurutamise intensiivsus seletub suurelt osalt suhteliselt lihtsate nõuetega, mida nad esitavad keele enda automaatsele analüüsile.

Ent seesama asjaolu on ka IOS-ide puuduste allikas. Peamiseks puuduseks on põhimõtteliselt piiratud täpsus, millega süsteem suudab päringule vastata.

Selle üle, kas mingi tekst sisaldab informatsiooni päringus esitatud küsimuse kohta, otsustab süsteem selle põhjal, kas tekstis leiduvad mõisted, millest päring koosneb, või ei leidu. Ent niimoodi saab kindlaks teha ainult seda, et vaatlusaluses tekstis on juttu neist nähtustest, objektidest jms., mida vastavad mõisted kajastavad, mitte aga seda, mida täpselt nende kohta räägitakse, mida mille kohta väidetakse või eitatakse, lühidalt, mis seostesse nad on asetatud.

Selline süsteem ei erista näiteks fraase *ülema solvamine alluva poolt* ja *alluva solvamine ülema poolt*, *vanema kõrvalehoidumine lapsele alimentide maksmisest* ja *lapse kõrvalehoidumine vanemale alimentide maksmisest*, sest nende fraaside mõisteline koosseis on vastavalt sama.

Põhjus peitub asjaolus, et eespool kirjeldatud infokeeltes arvestatakse ainult üht tüüpi seoseid deskriptorite vahel, nimelt paradigmaatilisi seoseid, mis on fikseeritud aprioorselt ja kehtivad kogu vastava ainevalla ulatuses sõltumatult konkreetsetest tekstidest. Neis ei arvestata üldse seda, kuidas on omavahel seotud deskriptorid, mis esinevad koos ühes kindlas tekstis, eriti näiteks deskriptorid, mis esinevad koos igas üksikus lauses. Teisiti öeldes, neis ei tooda esile süntagmaatilisi seoseid deskriptorite vahel.

Püüetega võtta arvesse ka viimati mainitud seoseid ja kujutada tekstide tõlkeid infokeeltesse mitte lihtsalt deskriptorite loenditena, vaid seotud struktuuridena, mis annaksid tõlgitavate tekstide sisu täpsemini edasi, ongi seotud põhilised arengutendentsid automatiseeritud IOS-ide loomisel.[12]

Süntagmaatiliste seoste esiletoomine ja arvestamine automatiseeritud infootsisüsteemis on hoopis raskem ülesanne kui paradigmaatiliste seoste arvestamine. Viimased kehtivad, nagu öeldud, kogu vastava ainevalla ulatuses: *töötasu* ja *kuutöötasu*, *kriminaalkaristus* ja *vabadusekaotus* on soo-liigi-seoses, sõltumata seadusandlikest aktidest, kus üks või teine mõiste neist paaridest esineb, nad on seda «väljaspool» konkreetseid tekste. Seepärast on võimalik selliseid seoseid deskriptorite vahel ette ära näidata, fikseerida nad täiesti kindlalt; seda ülesannet täidavadki tesaurused. Ent süntagmaatilised seosed kahe deskriptori vahel võivad varieeruda ühest tekstist teise, ühest lausest teise, mistõttu neid ei saa iga üksiku deskriptori puhul eelnevalt fikseerida.

жений с ядерными конструкциями. Научно-техническая информация. 1974, сер. 2, nr. 1, lk. 22—26.

[12] Vt. Н. А. Стоколова, О тенденциях в области разработки информационно-поисковых языков. Исследования по математической лингвистике, математической логике и информационным языкам. Москва, 1972, lk. 160—199.

Nad tuleb kindlaks määrata tekstide infokeelde tõlkimise käigus. Automaatse indekseerimise korral tähendab see aga, et süntagmaatilised seosed deskriptorite vahel on tarvis tekstides automaatselt ära tunda. See ongi asjaolu, mis nende seoste arvestamise raskeks teeb: nende automaatne äratundmine nõuab võrdlemisi põhjalikku tekstide automaatset analüüsi, seejuures mitte ainult süntaktilist, vaid teatavas ulatuses ka semantilist analüüsi. Küsimus pole ju lihtsalt selles, et tuleb määrata süntaktilised vahekorrad sõnade vahel, vaid on tarvis ära tunda ka sisulised seosed, mis seovad sõnadega väljendatavaid mõisteid. Näiteks: on vaja kindlaks teha, kuidas vormiliselt on seotud sõnad fraasis *ülema solvamine alluva poolt*, ja ühtlasi ka ära tunda, et samades sisulistes seostes on need sõnad (või mõisted) näiteks lauses *alluv on solvanud ülemat*.

Kuid see analüüs ei pea siiski olema kaugeltki nii põhjalik kui masintõlkimise puhul ning vastavatelt algoritmidelt ei nõuta 100-protsendilist töökindlust.

Esiteks ei ole põhimõtteliselt midagi katki, kui süsteem ei suuda teksti analüüsides üht või teist seost kindlaks teha. See tähendab ainult, et päringu peale, mille puhul vastav seos on relevantne, annab süsteem sellevõrra vähem täpse vastuse, kuid see vastus on siiski täpsem kui süsteemil, mis opereerib ainult deskriptorite loenditega.

Teiseks ei ole tarvis, et süsteem suudaks analüüsida kogu teksti, kõiki konstruktsioone, mis selles üldse ette tulevad. Võib lähtuda eeldusest, et ka süntaktilis-semantilisel tasandil kehtib seaduspärasus, mis on omane keele teistele tasanditele. Vaadeldaval juhul avaldub niisugune seaduspärasus selles, et on olemas piiratud hulk suhteliselt lihtsaid konstruktsioone, mida kasutatakse valdaval enamikul kordadel ja mille abil seega antakse edasi põhiline osa informatsioonist. Nende konstruktsioonide analüüsimisele tulebki tähelepanu koondada, kuna suure hulga keerulisi, kuid suhteliselt harva kasutatavaid konstruktsioone võib esialgu tähele panemata jätta.

Sellest eeldusest lähtudes on koostatud süntaktilise analüüsi programmid näiteks IOS-is «Pusto—Nepusto», ühes tuntumas Nõukogude Liidus väljatöötatud süsteemis, milles arvestatakse süntagmaatilisi seoseid.[13]

Teiselt poolt ei ole vaja kindlaks teha kõiki tekstides leiduvaid sisulisi vahekordi. Sõltuvalt IOS-i temaatikast valitakse teatav hulk vahekordi, mis on sellele temaatikale spetsiifilised, ja suunatakse jõupingutused just nende väljaselgitamisele. Kui näiteks keemias on olulised vahekorrad, mis koonduvad reaktsiooni mõiste ümber (reaktsioon — lähteained, reaktsioon — tulemus), siis näiteks seadusandlikes tekstides on relevantsed vahekorrad, mis on seotud tegevuse mõistega (tegevus — tegija; tegevus — tagajärg; tegevus — vahend jne.).

On püütud koostada ka universaalseid sisuliste seoste loendeid, mis eeldatakse olevat relevantsed mis tahes alal. Seda eesmärki silmas pidades on loodud näiteks SYNTOL (akronüüm nimetusest *Syntagmatic Organization Language*),[14] üks tuntumaid süntagmaatiliste seostega infokeeli. Selles eristatakse kõigepealt nelja seost: 1) predikatiivne, 2) konsekutiivne, 3) koordinatiivne ja 4) assotsiatiivne seos. Neist kaht seost võib eriliste operaatoritega täpsustada: assotsiatiiv-

---

[13] И. С. Добронравов, Д. Г. Лахути, Г. А. Лесскис, Об одном подходе к разработке автоматизированной ИПС с грамматикой. Научно-техническая информация. 1973, сер. 2, nr. 6, lk. 17—19.

[14] R. C. Cros, J. C. Gardin, F. Levy, SYNTOL — Syntagmatic Organization Language. Paris, 1964.

set võib konkretiseerida vahendi, koha, eesmärgi ja tunnuse operaatoritega, koordinatiivset võrdluse, identifitseerimise ja diferentseerimise operaatoritega.

Kuid oma abstraktsuse ja universaalsuse tõttu on niisuguseid seoseid konkreetsetes tekstides juba raske automaatselt kindlaks teha. Nad nõuavad väga häid süntaktilise analüüsi programme, mistõttu praktiliselt töötavates IOS-ides pole sellist universaalset infokeelt seni mõeldav kasutada.

Ülaltoodus avaldub selgesti see, mille poolest infotöötlus on sobivam ja mugavam keele automaatse analüüsi meetodite väljatöötamiseks kui näiteks masintõlkimine. Masintõlkesüsteemi loomine nõuab otsekohe maksimaalseid tulemusi, nii lihtsate kui keeruliste probleemide lahendamist — enne süsteem tervikuna tööle ei hakka. Infosüsteemi raames võib aga edasi liikuda järk-järgult, lahendades algul kõige lihtsamaid probleeme, seejärel üha keerulisemaid, kusjuures süsteem võib hakata tööle juba kõige lihtsamate analüüsimeetodite olemasolu korral.

Teiselt poolt võib infosüsteemi raames hakata otsekohe lahendama ka kõige keerulisemaid keeleprobleeme, mis on seotud tekstide semantilise analüüsiga. Kuid siingi pole vaja lahendada kõiki probleeme korraga ja täies ulatuses. Võib valida mingi kitsa temaatikaga tekstid (see aga lihtsustab semantilist analüüsi tunduvalt), edasi võib valida ainult teatava kitsama ringi keelenähtusi ning üksnes nendes fikseeritud piirides hakata sügavuti tungima, kuna ülejäänud keelenähtuste puhul võib jääda tavalise «pindmise» analüüsi tasemele.

Semantilise analüüsi probleemid viivad meid aga juba teise süsteemitüübi — semantiliste infotöötlussüsteemide juurde. Kuid enne kui neid vaatlema hakkame, esitame siin mõnede automatiseeritud infootsisüsteemide kirjeldused, et anda seni kirjeldatud süsteemitüübist terviklikumat pilti.

MÕNED INFOOTSISÜSTEEMID. Üks tuntumaid infootsisüsteeme maailmas on SMART (= *Salton's Magical Automatic Retriever of Texts*), mis loodi Harvardis ajavahemikus 1962—1965 tuntud infotöötlusspetsialisti G. Saltoni juhtimisel.[15] SMART on täielikult automatiseeritud süsteem, mis töötleb loomulikus keeles dokumente ja päringuid ilma igasuguse eelredigeerimiseta. Tekstide töötlemiseks võidakse süsteemis kasutada mitutsada erinevat analüüsimeetodit. Nii suur analüüsimeetodite hulk on seletatav asjaoluga, et SMART pole mõeldud mitte lihtsalt tekstide automaatseks otsimiseks, vaid omamoodi eksperimentaalseks baasiks mitmesuguste tekstitöötlusmeetodite võrdlemisel. Need meetodid võimaldavad määrata ka sõnade tähenduslikku lähedust statistilise analüüsi teel, samuti sooritada sõnaühendite automaatset süntaktilist analüüsi jms. Süsteemi koosseisu kuulub deskriptorite sõnastik (tesaurus), milles on esile toodud sünonüümsus, sooliigi-seosed ja assotsiatiivsed seosed (määratud eelmainitud statistilise analüüsi meetoditega), samuti kuulub sellesse tüüpiliste sõnaühendite sõnastik.

Kõiki neid vahendeid saab kasutada niimoodi, et päringud, millele on tulnud ebarahuldavad vastused, töödeldakse uuesti, kuid muudetud tingimustes ja teiste meetoditega; uut vastust analüüsitakse ja vajaduse korral muudetakse jällegi päringut ning korratakse analüüsiprotseduuri, kuni saadakse vastus nõutud kujul ja ulatuses.

---

[15] Põhjaliku ülevaate sellest süsteemist annab: Г. Сэлтон, Автоматическая обработка, хранение и поиск информации.

Nõukogude Liidus loodud IOS-idest on ilmselt tuntuim süsteem «Pusto—Nepusto», mis on välja töötatud Üleliidulises Teadus- ja Tehnikainformatsiooni Instituudis (ВИНИТИ). Õieti pole see üks süsteem, vaid süsteemide sari, millest on realiseeritud mitu varianti. Kirjeldame järgnevalt süsteemi «Pusto—Nepusto-2», mis on kõnealusest sarjast loodud viimasena.[16]

Tööd süsteemi «Pusto—Nepusto-2» kallal algasid 1965. a. Süsteem on mõeldud ühelt poolt praktiliseks ekspluateerimiseks elektrotehnika infokeskuses, teiselt poolt aga (niisamuti nagu ülalkirjeldatud SMART) on ta eksperimentaalseks baasiks automatiseeritud infootsisüsteemides kasutatavate keele automaattöötluse meetodite uurimisel.

Süsteemi massiivi moodustavad elektrotehnikaalasest referaatajakirjast võetud referaadid. Süsteem on täielikult automatiseeritud; kasutatakse deskriptorite sõnastikku, milles on fikseeritud sünonüümsus ja soo-liigi-seosed. Viimastel aastatel on tekstide analüüsimise meetodeid tunduvalt edasi arendatud, eelkõige selleks, et infokeeles kajastada ka süntagmaatilisi seoseid deskriptorite vahel. Selleks on koostatud ja süsteemi koosseisu lülitatud spetsiaalne automaatse süntaktilise analüüsi programm, millest oli veidi juttu juba eespool.[17]

Et see on mitmes mõttes tüüpiline IOS-ides kasutatav süntaktilise analüüsi programm, siis iseloomustame teda veidi lähemalt.

Nagu eespool öeldud, on programm mõeldud vaid piiratud arvu lihtsamate konstruktsioonide analüüsimiseks. Need valiti välja elektrotehnikat käsitlevate referaatide eelneval läbitöötamisel (umbes 70 eri süntagmat). Samuti on piiratud sõnastik, millega programm opereerib (esialgses variandis kuulus sellesse umbes 300 täistähenduslikku ja 90 abisõna).

Süntaktilise analüüsi protseduur ise on võrdlemisi tüüpiline. Tekst jaotatakse lauseteks, mis seejärel segmenteeritakse, s. t. eraldatakse nendes välja vormiliselt piiritletud lõigud — pea- ja kõrvallaused ning nendes omakorda süntaktilised rühmad. Järgneb segmendisisene analüüs, mille käigus püütakse konstrueerida iga segmendi süntaktiline struktuur sõltuvusvahekordade järgi. Selleks toetutakse sõnastikule, kus iga sõna puhul on osutatud, missugustes süntagmades see võib olla pea-, missugustes sõltuvaks elemendiks. Lause terviklik struktuur saadakse järgmise etapi, nn. segmentidevahelise analüüsi tulemusena, kus segmendisisese analüüsi teel saadud struktuurid ühendatakse.

Lõpuks iseloomustame paari sõnaga ka automatiseeritud IOS-i, mille loomisega tegeldakse Tartus TRÜ kriminoloogia laboratooriumis (selle väljatöötamisest võtab osa ka käesoleva kirjutise autor), — seadusandlike tekstide automatiseeritud otsimise süsteemi JURIOS.[18] Võib märkida, et see on esimene suuremõõtmeline praktiliseks kasutamiseks mõeldud IOS Eesti NSV-s. Et süsteem pole veel lõplikult valmis, ei saa rääkida sellest, missugune ta on, vaid sellest, missugune ta tuleb. Ent me ei räägi ainult projektide põhjal. Suurem osa tõid on lõpule viidud ja veel käesoleval aastal algab süsteemi katsetamine.

---

[16] Д. Г. Лахути, Поисковая система «Пусто—Непусто-2». Труды III Всесоюзной конференции по информационно-поисковым системам и автоматизированной обработке научно-технической информации. Т. I. Москва, 1967, lk. 101—106. Samas on teisi artikleid selle süsteemi mitmesuguste aspektide kohta.

[17] И. С. Добронравов, Д. Г. Лахути, Г. А. Лесскис, Об одном подходе к разработке автоматизированной ИПС с грамматикой.

[18] В. Ю. Раудсалу, И. А. Ребане, И. Я. Сильдмяэ, О создании автоматизированной системы юридической информации. «Советское государство и право» 1974, nr. 5, lk. 28—36.

JURIOS-e massiivi moodustavad eestikeelsed seadusandlikud tekstid: seadused, Ministrite Nõukogu määrused, korraldused jms.

Seadusandlikud tekstid on mitmes mõttes sobiv materjal IOS-i loomiseks ja üldse automaatseks keeleliseks töötlemiseks. Juriidilised normid peavad ühtekokku moodustama kindla süsteemi, seda vähemalt iga õigusharu piires (kriminaalseadusandlus, tsiviilseadusandlus jne.). Teiseks iseloomustavad norme kindlad loogilised iseärasused, kindel loogiline struktuur ja loogilised vahekorrad üksiknormide vahel.[19]

Kuid seadusandlikel tekstidel on ka iseärasusi, mis kui mitte otseselt ei raskenda IOS-i loomist, siis kitsendavad tunduvalt võimalike lahenduste hulka. Kõigepealt ei ole seaduste puhul mõeldav opereerida sisukokkuvõtetega, referaatidega, nii nagu seda võib teha näiteks teadusliku kirjanduse puhul. Iga seadus kehtib mitte ainult täpselt selles sisus, vaid ka täpselt selles sõnastuses, nagu seadusandlik organ on ta andnud. Seetõttu tuleb süsteemi mällu viia seaduste täielikud tekstid ja opereerida nendega.

Teiseks haaravad seadusandlikud tekstid väga laia ja mitmekesist ainestikku. Õigus reguleerib väga erinevaid elualasid ning kõigi nende alade sõnavara figureerib seadusandlikes aktides. See komplitseerib tunduvalt infokeele loomist, eriti deskriptoritevaheliste seoste määramist.

JURIOS on mõeldud täielikult automatiseeritud süsteemina; nii süsteemi mällu viidavad tekstid kui ka päringud tõlgitakse infokeelde automaatselt. Süsteemi tesauruses, millesse kuulub üle 10 000 deskriptori, arvestatakse lisaks sünonüümsusele ja soo-liigi-seosele veel kuut liiki seoseid.[20] Neist võib eriti mainida nn. funktsionaalset seost, mis haarab vahekordi «tegija — tegevus» ning «tegevus — tagajärg», ja nn. juriidilist alluvusvahekorda, mis fikseerib asutuste, organisatsioonide jm. subordinatsiooni.

Esialgu, praktiliseks kasutamiseks mõeldud variandis ei arvestata JURIOS-es süntagmaatilisi seoseid; ka ülalmainitud funktsionaalset seost kasutatakse paradigmaatilise seosena, see fikseeritakse ainult seal, kus ta loogiliselt kehtib vastavate mõistete vahel: *kurjategija — kuritegu, tapmine — surm* jms.

Küll aga on juba käsil tööd selleks, et muuta süsteem kakskeelseks: et selle massiivi võiksid kuuluda lisaks eestikeelsetele ka venekeelsed seadusandlikud aktid (suurt osa üleliiduliselt kehtivaid seadusi ei tõlgita eesti keelde, kuid nad kehtivad meil) ja et ka päringu võiks esitada emmas-kummas keeles.

(Järgneb)

---

[19] Normide loogilistele omadustele on ammu tähelepanu pööranud loogikud. On välja kujunenud eriline loogikaharu — deontiline loogika, mis tegeleb normide ja normatiivsete arutluskäikude formaalloogilise analüüsiga. Ülevaate saamiseks sellest võib soovitada: А. А. Ивин, Логика норм. Москва, 1973.

[20] И. Г. Кулль, И. Я. Сильдмяэ, А. К. Хелемяэ, Х. Я. Ыйм, О разработке тезауруса юридических терминов для информационно-поисковой системы. Правовая кибернетика. Москва, 1973, lk. 54—62.

Keele automaattöötlus ja automatiseeritud infosüsteemid

(Algus „Keeles ja Kirjanduses" nr. 8)

HALDUR ÕIM

Semantilised infotöötlussüsteemid

Eelnenus kirjeldasime infootsisüsteeme — keele automaattöötluse süsteeme, mis on mõeldud praktiliseks ekspluateerimiseks. Järgnevalt kirjeldatavate süsteemide puhul ei peeta silmas otsest kasutamist praktikas; nad on mõeldud peamiselt uute, täiuslikumate infotöötlusmeetodite otsimiseks ja teoreetiliste probleemide läbitöötamiseks. Kuid neidki edasiviivaks jõuks on peale teoreetilise huvi ka suurel määral praktika surve.

Kui eespool kirjeldatud automatiseeritud infootsisüsteeme ei saa pidada täiuslikuks süsteemitüübiks omal alal, kui nad on pigem hädaabinõu, siis seda eelkõige nende lingvistilise külje suhtelise nõrkuse tõttu. Tee täiuslikumate süsteemide poole läheb paremate keele töötlemise meetodite kaudu. Seisukohta, et automatiseeritud infotöötlussüsteemides tuleb peatähelepanu pöörata nende «lingvistilise varustuse» täiendamisele, on väljendanud enamik selle ala spetsialiste; seda leiti ka VMAN-i liikmesriikide ühisel seminaril möödunud aastal.[21]

Põhilised probleemid, mis siin nõuavad lahendamist, on seotud semantikaga. Täiuslik süsteem on see, mis suudab ilma oluliste kadudeta kätte saada tekstidesse kätketud sisu ja selle sisulise informatsiooniga opereerida; tekstide sisu esiletoomine ja sobiva formaalse esitusviisi leidmine on aga semantika ülesanne. Seepärast on niisuguse infokeele põhimõtete väljatöötamine, mis semantilisel tasemel oleks põhimõtteliselt võrdne loomuliku keelega, tunnistatud keele automaattöötluses (ja infolingvistikas) esmase tähtsusega ülesandeks: «On küpsenud vajadus niisuguse infokeele loomise järele, mis sarnaselt loomuliku keelega oleks orienteeritud kogu selle andmehulga kajastamisele, mis iseloomustab «naiivset maailmapilti». Olulisim ideaalne nõue sellisele infokeelele, mille loomine on informaatika ja lingvistilise semantika ühine ülesanne, on kõigi loomulikus keeles eristatavate tähendussuhete algoritmiline äratuntavus.»[22]

Lingvistiline semantika on viimastel aastatel jõudsasti edasi arenenud ning selle tulemuste olulisus keele automaattöötluse seisukohalt on kaheldamatu. Kuid puhtlingvistiline semantika haarab ainult ühe poole semantikaprobleemistikust, mille lahendamine on vajalik eesmärgi saavutamiseks; need on keelelise tähenduse olemuse, tähenduse ja vormi vahekorra ning (osalt) tähenduste formaalse esitamise probleemid. Lingvistiline semantika ei puuduta aga probleemistiku teist

[21] И. С. Кравцова, Семинар стран — членов СЭВ «Автоматическая обработка текстов на естественных языках». Научно-техническая информация. 1973, сер. 2, nr. 6, lk. 49—50.

[22] С. И. Гиндин, Соотношение естественных и искусственных языков (научная конференция в Институте языкознания). Научно-техническая информация. 1973, сер. 2, nr. 6, lk. 25. Vt. samuti: Г. Э. Влэдуц, О соотношении естественных и искусственных языков. Семантические проблемы науки, терминологии и информатики. Т. I. Москва, 1971, lk. 59—63.

poolt: esiteks seda, kuidas tekstides varjul olev tähendus automaatselt kätte saada (automaatse semantilise analüüsi probleem), ja teiseks, kuidas tähendustega infotöötlusprotsessis opereerida, s. t. milles need operatsioonid seisnevad, millest nad sõltuvad, kuidas neid formaalselt kirjeldada. See kõik tuleb lahendada infolingvistidel endil.[23]

Peamiseks mooduseks viimati loetletud probleemidele sobivate lahenduste otsimisel on eksperimentaalsete infotöötlussüsteemide loomine. Infotöötluse eeliseks (võrreldes näiteks masintõlkega) on seejuures, nagu eespool juba osutasime, asjaolu, et igas üksiksüsteemis ei ole vaja korraga läbi töötada kõiki probleeme, mille lahendamist semantiline infotöötlus nõuab. Iga süsteemi puhul võib tähelepanu koondada mingite probleemide ringile.

Semantiliste süsteemide põhierinevus eespool kirjeldatud süsteemidest tuleneb nende põhimõtteliselt erinevast funktsioonist informatsiooni vahendamisel: süsteemile esitatud küsimuse peale ei viita nad allikatele, vaid annavad sisulise vastuse. See tähendab, semantiliste süsteemide mälus pole talletatud mitte deskriptorite abil liigitatud dokumendid, vaid faktid ise — nende dokumentide faktsisu — ja vastuseks päringule ei väljasta süsteem mitte dokumente ega viiteid dokumentidele, vaid nõutud fakte. Seetõttu nimetatakse vaadeldavaid süsteeme sageli faktograafilisteks ehk f a k t i o t s i s ü s t e e m i d e k s, vastandades neid dokumendiotsisüsteemidele. Oluline on seejuures, et faktograafilise süsteemi mällu salvestatu ei kujuta endast lihtsalt isoleeritud faktide kogumit. Süsteem peab suutma opereerida mälus leiduva informatsiooniga selliselt, et ta iga päringu puhul väljastaks kogu informatsiooni, mis küsimuse kohta tegelikult tema mälus sisaldub, ja mitte lihtsalt need andmed, mis talle vahetult on teatatud. Näiteks kui süsteemile on teatatud faktid «Magnetlindid on välismälu liik» ja «Arvutis «Minsk 22» kasutatakse magnetlinte», siis peab süsteem suutma vastata jaatavalt küsimusele «Kas arvutis «Minsk 22» kasutatakse välismälu?» (näide on võetud G. Iljini jt. artiklist, vt. viide 23). Või veel ilmekam, ehkki üpris lihtne näide: kui süsteemile on teatatud, et inimese käel on viis sõrme ja et inimesel on kaks kätt, peab süsteem suutma iseseisvalt vastata küsimusele, mitu sõrme on inimesel (näide B. Raphaeli süsteemist SIR).

Süsteemi niisuguse taseme saavutamiseks on vaja, et kõigi sõnade, samuti sisulist informatsiooni kandvate grammatiliste elementide ja süntaktiliste konstruktsioonide tähendused oleksid esitatud (defineeritud) semantiliste elementaarüksuste ja -suhete kaudu. Viimaste loogikaomadused ja omavaheliste kombinatsioonide võimalused peavad olema fikseeritud reeglite abil. Semantilised üksused ja suhted, nende omadusi kirjeldavad ning nende võimalikke kombinatsioone määravad reeglid moodustavad s e m a n t i l i s e k e e l e. See vastab eespool kirjeldatud IOS-ide infokeeltele. Semantiliste süsteemide üleolek IOS-idest võrsubki peamiselt nendes kasutatavate semantiliste keelte suuremast täielikkusest IOS-ide infokeeltega võrreldes. Semantilise keele abil on esitatud kõik andmed süsteemi mälus, süsteemi «teadmised»; see, kuivõrd süsteem suudab talle vahetult teatatud informat-

[23] M. P a c k, A. W. P r a t t, The Function of Semantics in Automated Language Processing. Proceedings of the 1971 Symposion on Information Storage and Retrieval. Maryland, 1971, lk. 5—18; D. G. H a y s, Applied Computational Linguistics. Applications of Linguistics. Selected Papers of the Second Congress of Applied Linguistics. Cambridge, 1969. Cambridge, 1971, lk. 69—85; Г. М. Ильин, Б. М. Лейкина, Т. Н. Никитина, С. И. Откупщикова, С. Я. Фитиалов, Лингвистический подход к задаче построения информационной системы. Информационные вопросы семиотики, лингвистики и автоматического перевода. Вып. 2. ВИНИТИ. Москва, 1971, lk. 4—13.

sioonist tuletada mitmesuguseid kaudseid andmeid, implitsiitset informatsiooni, sõltub sellest, kuivõrd süsteemis kasutatava semantilise keele reeglid sellist andmetega manipuleerimist võimaldavad.

Kõik süsteemi mällu viidavad tekstid tuleb tõlkida semantilisse keelde. Teiselt poolt peab süsteem tõlkima semantilisest keelest loomulikku keelde andmed, mis esitatakse vastusena päringule. Seejuures on tekstide automaatse analüüsi programmidele niisugustes süsteemides iseloomulik semantika aktiivne kasutamine, eriti näiteks selleks, et piirata süntaktilise analüüsi algoritmi poolt pakutavate võimalike analüüsivariantide hulka (võimalike variantide rohkus on automaatses süntaktilises analüüsis ikka tüli teinud).

Mis puutub äsja kirjeldatud üldpõhimõtetele vastavatesse tegelikesse süsteemidesse, siis hakati neid hoogsamalt looma 1960-ndate aastate keskpaiku ja eelkõige USA-s, kus neid nimetatakse «küsimustevastuste süsteemideks» (*question-answering systems*).[24] Esimestele süsteemidele oli iseloomulik opereerimine mingi hästi kitsapiirilise ja kergemini formaliseeritava ainestikuga (nagu näiteks inimeste sugulusvahekordi puudutav informatsioon), samuti lihtsate infotöötlusmeetodite kasutamine. Ent järk-järgult on infotöötlusmeetodid täienenud ja ka keele analüüsimise võime kasvanud.

1960-ndate aastate keskel loodud süsteemidest on saanud eriti tuntuks B. Raphaeli süsteem SIR (=*Semantic Information Retriever*).[25] Eriti hoolikalt on välja töötatud selle semantiline keel ja loogikavahendid, mida süsteem saab kasutada oma «arutluskäikudes». Nii see keel kui ka need vahendid on saanud eeskujuks paljudele hiljem loodud süsteemidele. Tekstide analüüsimises on SIR-i võimed aga võrdlemisi keskpärased; vastuvõetavate konstruktsioonide hulk on rangelt piiratud.

Üheks paremaks seni loodud semantiliseks süsteemiks võib pidada R. M. Schwarczi, J. F. Burgeri ja R. F. Simmonsi süsteemi Protosynthex III.[26] Selles on püütud samavõrra hoolikalt välja töötada kõik kolm aspekti, mis niisuguse süsteemi ilme määravad: semantiline keel tekstide sisu esitamiseks; semantilised infotöötlusmeetodid; tekstide analüüsimise ja sünteesimise algoritmid. Seetõttu tasub seda süsteemi veidi lähemalt kirjeldada.

Andmed Protosynthex III mälus on esitatud sündmusteks (*event*) nimetatavate elementidena, milledel on üldkuju XRY, kus R kujutab endast mingit suhet ja X ning Y on üksused, mis võivad olla semantiliselt elementaarsed mõisted, aga ka kui tahes komplekssed. Lisaks on mõisted omavahel korrastatud hierarhiatesse vastavalt nende loogikaomadustele. Eriline koht on nn. semantilistel sündmusvormidel, kus teatud üldmõistete kaudu on näidatud, missugused mõistekombinatsioonid on semantiliselt võimalikud. Semantilisi sündmusvorme kasutab süsteem tekstide analüüsimisel aktiivselt selleks, et kõrvale heita sisuliselt mõttetuid analüüsivariante.

Samuti on süsteemis hoolikalt koostatud reeglid mälus leiduva

---

[24] Võrdlemisi täieliku pildi semantiliste süsteemide loomise etappidest ja olulisematest süsteemidest võib saada kahest R. F. Simmonsi ülevaateartiklist: R. F. Simmons, Answering English Questions by Computer: a Survey. "Communications of ACM" 1965, vol. 8, nr. 1, lk. 53—69; R. F. Simmons, Natural Language Question-Answering Systems 1969. "Communications of ACM" 1970, vol. 13, nr. 1, lk. 3—20. Mitmed tuntud 1960-ndatel aastatel loodud süsteemid on koondatud kogumikku: Semantic Information Processing. Ed. M. Minsky. Cambridge (Mass.), 1968.

[25] B. Raphael, SIR: a Computer Program for Semantic Information Retrieval. Semantic Information Processing, lk. 33—134.

[26] R. M. Schwarcz, J. F. Burger, R. F. Simmons, A Deductive Question-Answerer for Natural Language Inference. "Communications of ACM" 1970, vol. 13, nr. 3, lk. 167—183.

materjaliga opereerimiseks, eriti nende järelduste tuletamiseks, mis on vajalikud küsimustele vastuste leidmiseks. Süsteemi sellekohastest võimetest võib anda pildi järgmine näide.

Süsteemile on esitatud tekst: «On ahv. On kast. On banaanid. Banaanid on kõrgemal kui ahv. Kast on kõrgusega ese. Ahv toimetab kasti banaanide juurde. Ahv seisab kastil. Ahv sirutab käe banaanide poole.»

Küsimus: «Kas ahv saab banaanid kätte?»

Sellele küsimusele vastamiseks teeb süsteem läbi järgmise arutluskäigu, mille ta koos vastusega väljastab: «Ahv saab banaanid kätte, kui ta ulatub nendeni, ja ta võib ulatuda nendeni, kui ta seisab mingil kõrgusega esemel nende all ja sirutab käe nende poole. Ahv seisab kastil, mis on kõrgusega ese, ja niisiis kui kast on banaanide all, ulatub ahv nendeni. Kast on banaanide all, kui see on madalamal kui banaanid, mida ta on, ja on ka banaanide juures; kast on banaanide juures, kui keegi on selle sinna toimetanud — mida ahv tõepoolest on teinud. Seega ahv ulatub banaanideni.» Selle järelduse süsteem esitabki vastuseks.

Mis puutub tekstide analüüsisse, siis selles on siiski ka Protosynthexi võimed tublisti piiratud. Ta tuleb toime ainult suhteliselt lihtsate lausete ja tekstidega (nagu võib näha ka eeltoodud näitest).

Meil NSV Liidus hakati semantiliste süsteemide loomisega tegelema alles hiljuti, seepärast on küll üsna rohkesti projekte, kuid vähe valminud ja funktsioneerivaid süsteeme.

Esimeseks sellelaadseks süsteemiks meie maal võib pidada Kiievis E. Skorohhodko juhtimisel loodud süsteemi BIT, mille ülesandeks on arvutustehnika kirjanduse töötlemine ja otsimine.[27] Semantilise süsteemi määratluse alla sobib BIT siiski ainult pooliti. Süsteemi mällu viidavad tekstid ja ka päringud analüüsitakse küll süntaktiliselt ja semantiliselt ning esitatakse erilises semantilises keeles, nn. RX-koodis, kuid vastusena väljastab süsteem ainult viited allikatele, kust andmeid leida. Tõsi küll, süsteem võib viited varustada lihtsustatud annotatsioonidega, mis koosnevad viidatava dokumendi sisu kirjeldavatest lausetest. Seevastu on aga BIT-iga tehtud eksperimendid andnud materjali rohketeks infolingvistilisteks uurimusteks.[28]

Kõige selgemalt on meil semantilise süsteemi loomise idee tõstnud vahest teadlased Leningradi RÜ matemaatilise lingvistika laboratooriumist.[29] On välja töötatud süsteemi üldised põhimõtted ja tehtud tegelikke keeleuurimusi süsteemi loomise raames.[30] Töötava süsteemini pole seni veel jõutud.

Tulemuste poolest on aga kindlasti kõige paremini tuntud Moskva võõrkeelteinstituudi masintõlkelaboris tehtavad tööd. Neist puudutavad meie teemat eriti N. Leontjeva uurimused, mille eesmärk on tekstide tähenduse esitamiseks vajaliku formaalse keele loomine ja

---

[27] Информационно-поисковая система «БИТ». Киев, 1968.

[28] Vt. näit. artikleid kogumikes: Семантические проблемы автоматизации информационного поиска. Киев, 1971; Лингвистические проблемы автоматизации информационного поиска. Киев, 1973; Математическая лингвистика. Ежегодник по структурной, прикладной и математической лингвистике. Киев, 1973.

[29] Г. М. Ильин, Б. М. Лейкина, Т. Н. Никитина, М. И. Откупщикова, С. Я. Фитиалов, Модель семантики текста и система «запрос — ответ» (к постановке задачи). Научно-техническая информация 1969, сер. 2, nr. 1, lk. 10—14.

[30] Информационные вопросы семиотики, лингвистики и автоматического перевода. Вып. 2. Москва, 1971; Лингвистические проблемы функционального моделирования речевой деятельности. Вып. 1. Ленинград, 1973.

sellekohaste analüüsiprogrammide väljatöötamine infosüsteemi raames.[31]

N. Leontjeva eristab teksti informatsioonilist esitust selle semantilisest esitusest. Viimane koosneb teksti moodustavate üksiklausete semantilistest esitustest. Teksti informatsiooniline esitus on aga tervik, mis saadakse üksiklausete semantiliste esituste ühendamisel vastavalt lausete vahel kehtivatele seostele ja teksti terviklikule struktuurile. Viimasena viidatud artiklis kirjeldatakse teksti semantilise esituse koostamist.

Semantilises infokeeles, mida Leontjeva kasutab, eristatakse nn. terme (õieti deskriptoreid) ja kahekohalisi semantilisi seoseid. Termidel on enamasti kompleksne struktuur, mis esitatakse semantiliste tunnuste ja semantiliste seoste abil. Infokeele variandis, mida kirjeldatud eksperimendis kasutati, on näiteks 13 semantilist tunnust («tegu», «ese», «materjal», «asutus» jms.) ning 26 semantilist seost («seotud», «kuuluvus», «osa», «lokalisatsioon» jne.). Semantiliste tunnuste hulk sõltub konkreetsest ainestikust, mille esitamiseks infokeel on mõeldud, semantilised seosed on aga suhteliselt universaalsed: mitmesuguse sisuga tekstide kirjeldamiseks piisab Leontjeva arvates 40—100 seosest.

Olulist osa süsteemis lisaks otsestele analüüsiprogrammidele mängib semantiline sõnastik, milles sisaldub üksikasjalik informatsioon iga sõna kohta, mida süsteem tekstis kohtab: sõna semantiline kirje; andmed selle kohta, missuguste semantiliste seoste liikmeks see sõna saab olla ja missugused teised üksused saavad sel juhul täita seose teist kohta jne. Lausete analüüs tugineb oluliselt sõnastikule, seetõttu pole ette nähtud ka eraldi süntaktilise analüüsi tasandit, sest süntaktiline ja semantiline analüüs toimuvad korraga.

Lausete semantilised esitused kujutavad endast sõltuvuspuid, mille elementideks on termid (või semantilised tunnused) ja semantilised seosed. Teksti informatsioonilises esituses ühendatakse lausete semantilised esitused omakorda semantiliste seoste abil.

Süsteem on mõeldud faktograafiliseks infootsinguks, aga samuti mitmesugusteks muudeks operatsioonideks, mida võimaldab tekstide kokkusurutud esitus süsteemi mälus, näiteks referaatide ja annotatsioonide koostamiseks.

PERSPEKTIIVIDEST. Nagu ütlesime, on semantiliste süsteemide loomise peaeesmärk täiuslikumate analüüsi- ja infotöötlusmeetodite loomine. Eelkirjeldatud süsteemid esindavad niisiis infosüsteemi tüüpi, mis lähemas või kaugemas tulevikus peaks jõudma tegelikku kasutusse ja järk-järgult asendama praegused IOS-id.

Kuid ka neid süsteeme, mida siin kirjeldasime, ei saa veel kaugeltki nimetada täiuslikeks keele automaattöötluse süsteemideks. Nagu lugeja võis tähele panna, pidime igaühe puhul neist märkima, et süsteemil on teksti analüüsimise võimed piiratud mingi kitsama ringi lause- ja konstruktsioonitüüpidega. Osaliselt on niisugused kitsendused seletatavad sellega, et süsteemi loojad pole soovinud kõigisse keele analüüsimise üksikasjadesse süveneda, olles rohkem huvitatud süsteemi loogikavõimete arendamisest (nii on see näiteks SIR-i ja Protosynthexi puhul). Aga isegi ülidetailsete analüüsiprogrammide lisamine ei muudaks eelkirjeldatud süsteeme iseenesest sellisteks, mis suudaksid tekste täielikult mõista ja opereerida nendega nagu inimesed.

---

[31] Н. Н. Леонтьева, Создание информационного языка на базе семантического анализа текста. Научно-техническая информация 1971, сер. 2, nr. 8, lk. 8—15; Н. Н. Леонтьева, Е. В. Урысон, Алгоритм построения информационной записи текста. I этап. Научно-техническая информация. 1973, сер. 2, nr. 12, lk. 3—13.

Rõhutame seejuures, et jutt on just tekstide ja mitte üksiklausete mõistmisest. Kirjutada programme, mis suudaksid detailselt kirjeldada üksiklausete kõikvõimalikke tähendusvariante, ei ole nii väga keeruline. Kuid tekst ei ole lihtsalt lausete summa. Mis tahes teksti mõistmiseks kasutab inimene suurel hulgal sellist informatsiooni, mis otseselt pole selles tekstis antud ja pole seepärast ka avastatav kui tahes peene analüüsiga. Selleks et õpetada infotöötlussüsteemi mõistma tekste niisamuti nagu inimene, on vaja õpetada teda niisamuti opereerima tekstivälise informatsiooniga — ja eelkõige on muidugi vaja süsteem sellise informatsiooniga varustada.

On osutatud vähemalt kahte liiki informatsioonile, milleta ei ole mõeldav tekstide rahuldav mõistmine. Esiteks detailsed teadmised maailmast, millest tekstides kõneldakse. Tavalised tekstid, mis inimene on kirjutanud inimese jaoks, eeldavad väga suurel hulgal selliseid teadmisi. Näiteks mingit sündmust kirjeldades ei hakka me kunagi kõiki selle detaile üles lugema, vaid eeldame, et kuulaja või lugeja mõtleb need juurde; ühtlasi on üksiklausete seoseid tekstis võimalik mõista ainult nende juurdemõeldavate detailide kontekstis. Mõnesid jooni niisuguse «maailmamudeli» olemasolust — ja vajadusest selle järele — võisime täheldada juba eespool toodud Protosynthexi arutluses ahvist, kastist ja banaanidest: on vaja võrdlemisi üksikasjalikke teadmisi, missugune olend on ahv, mida kujutab endast kast, mida kujutavad endast ruumisuhted ja kuidas nad on seotud selliste tegevustega nagu «liigutama», «sirutama» või «ulatuma», selleks et mõista tekstis kirjeldatud situatsiooni.

Teiseks on leitud, et tekstide mõistmiseks on peale loogika vaja kasutada ka «praktilisi» arutluskäike, s. o. opereerida järeldustega, mis ei pruugi olla loogika mõttes korrektsed, kuid mis on tegelikkuses põhjendatud. Niisugune on näiteks järgmine arutlus: A teeb X; X põhjustab (tavaliselt) Y; järelikult A tahab Y. On täiesti loomulikud sellisedki tekstid, kus on lihtsalt teatatud «A teeb X», ja eelduse najal, et lugeja teab, et X põhjustab Y, on ühtlasi eeldatud, et lugejal on pärast eeltoodud teadet ka teada, et A tahab Y.

Nii ühe kui ka teise nõude rahuldamine viib õieti juba «puhaste» keele automaattöötluse süsteemide juurest teist tüüpi süsteemide — tehisintellektisüsteemide — juurde. Viimastel aastatel ongi hakatud teadlikult seostama keele analüüsimise (sealhulgas masintõlke) meetodite täiustamist tehisintellektisüsteemide joonte ülevõtmisega.[32]

Niisuguses asjade kulus ei ole iseenesest võttes midagi üllatavat. Keel on väga tihedalt seotud inimese tunnetus- ja mõtlemisprotsessidega. Need annavad keelele sisu. Ja selle sisu lahtimõtestamine nõuab omakorda mõtlemise ja tunnetuse seaduspärasuste arvestamist.

Kokkuvõttes ütleksime järgmist. Keele automaattöötlus võib tõhusalt kaasa aidata kitsamatel infootsingu ja -töötluse aladel. Põhiliseks keele automaattöötluse vormiks on seejuures automatiseeritud infootsisüsteemid.

Lähema tuleviku vormi esindavad semantilised infotöötlussüsteemid. Nende loomine on viimastel aastatel märgatavalt hoogustunud, sealhulgas ka meil NSV Liidus. Praegu suudavad semantilised infotöötlussüsteemid opereerida üksnes piiratud hulga tekstidega.

Süsteemid, mis suudaksid täielikult — inimese tasemel — mõista (ja tõlkida) loomuliku keele tekste, kuuluvad veel võrdlemisi kaugesse tulevikku.

[32] Näit. Y. Wilks, An Artificial Intelligence Approach to Machine Translation. Stanford Artificial Intelligence Project Memo, AIM-161. Stanford, 1972.