STATISTILISED MEETODID JA KEELETEADUS

J. Tuldava

Meie ajale on iseloomulik teaduste matematiseerumine, millest pole jäänud puutumata ka selline traditsiooniliselt mittematemaatiline teadus nagu keeleteadus. Matemaatiliste meetodite rakendamine keeleteaduses toimub tänapäeval kahes erinevas suunas, mis eeldab vastavalt mittekvantitatiivset ja kvantitatiivset lähenemist. Esimesel juhul toetutakse nn. diskreetsele matemaatikale, esmajoones matemaatilisele loogikale ja hulgateooriale. See suund uurib keele determineeritud omadusi. Teisel juhul võetakse aluseks tõenäosusteooria koos matemaatilise statistikaga ja informatsiooniteooriaga. Kvantitatiivne suund tegeleb keele mittedetermineeritud (statistiliste, tõenäosuslike) omadustega. Käesolevas töös lähtutakse matemaatilise lingvistika kvantitatiivsest suunast, s. o. statistilistest meetoditest keeleteaduses. Artikli esimeses osas vaadeldakse keelestatistika arengut ja rakendusvõimalusi tänapäeval. Eri alapeatükis antakse ülevaade keelestatistilisest uurimistööst Eestis. Artikli teises osas käsitletakse keelestatistika teoreetilisi aluseid ja nendest tuletuvaid praktilise töö põhimõtteid.

1. KEELESTATISTIKA ARENG JA RAKENDUSVÕIMALUSED TÄNAPÄEVAL

1.1. Keelestatistika ajaloost

Statistiline vaatlus keeleuurimistöös pole mingi uus nähtus. On andmeid selle kohta, et juba antiikajal ja keskajal tunti huvi tähtede ja sõnade esinemissageduse vastu tekstides. Pütaagorlased, kes arvasid, et "arvud valitsevad maailma", loendasid tähti sõnades ja uurisid helitute ning heliliste häälikute vaheldumist, omistades kvantitatiivsetele suhetele müstilisi omadusi. Keskajal loendati tähti ja sõnu piibli erinevates osades, juhendudes samuti müstilis-religioossetest ajenditest. 15. sajandist on aga tuntud Milaano elaniku Sicco Simonetta koostatud tähtede sagedustabelid ladina ja itaalia keele kohta, mille alusel tehti katset formaalselt eristada võrreldavaid keeli (Karlgren, 1968, 136).

Uuemal ajal huvituti keeleüksuste statistilisest uurimisest seoses praktiliste vajadustega, näit. katsetega koostada või dešifreerida salakirju. 19. sajandil teostati juba hulgaliselt keeleüksuste, eriti tähtede ja tähekombinatsioonide loendusi nii kirjutus- ja tüpograafiliste masinate otstarbekama ehituse kui ka stenograafiliste süsteemide väljatöötamise huvides. Esimesed selletaolised uurimused pärinevad 19. sajandi algusest ja on tehtud prantsuse keele põhjal. Statistilisi andmeid on ilmselt kasutatud ka morsetähestiku loomisel (näiteks on kõige sagedam inglise keele täht e edasi antud kõige lihtsama märgiga - ühe punktiga). Stenograafiliste süsteemide loomist pidas silmas ka esimese suure sagedussõnastiku looja sakslane F. Kaeding (1898).

Möödunud sajandi lõpul ja käesoleva sajandi esimesel veerandil huvitusid keele statistilisest uurimisest eriti psühholoogid, kes kasutasid statistilisi andmeid kõnetegevuse psühholoogilisel interpreteerimisel. Paljud nendest uurimustest pakuksid lingvistidele huvi ka tänapäeval. Kahekümnendate aastate töödest võib eriti esile tõsta psühholoogi A. Busemanni statistilisi vaatlusi laste kõnekeele kohta (Busemann, 1925). Mõningaid A. Busemanni keelestatis-

tilisi põhimõtteid (näit. sõnaliikide sageduste suhete kindlakstegemist) on hilisemal ajal hakatud laialdasemalt rakendama psühholingvistilistes uurimustes ja kvantitatiivses stilistikas.

Esimesed puhtlingvistilisi eesmärke taotlevad statistilised tööd ilmusid möödunud sajandil. Võib nimetada **E.Förstemanni kõrvutavat** uurimust kreeka, ladina ja saksa keele häälikute sageduste kohta (Förstemann, 1852) ja W.D. Whitney' inglise keele ja sanskriti häälikute kvantitatiivset analüüsi (Whitney, 1874). Ameeriklane T. Mendenhall tegi esimesena katset eristada individuaalseid stiile sõna- ja lausepikkuse ning nende statistilise jaotumuse alusel (Mendenhall, 1887). Sagedussõnastikke meenutavaid sõnaloendeid ja nn. konkordantse (näidiseid uuritavate sõnade kasutamise kohta antud teoses) on teadaolevatel andmetel koostatud juba möödunud sajandi alguses.

Lingvistilise suunitlusega statistilised uurimused olid möödunud sajandil ja ka käesoleva sajandi algul siiski võrdlemisi haruldased. Keeleteadlased huvitusid peamiselt keelenähtuste kvalitatiivsest analüüsist, lähtudes sisulistest, mitteformaalsetest kriteeriumidest. Kuid juba tol ajal juhtisid mõned tuntud keeleteadlased tähelepanu matemaatiliste, sealhulgas statistiliste meetodite vajalikkusele keeleuurimistöös. Nii näiteks kirjutas J. Baudouin de Courtenay 1901.a. vajadusest "sagedamini rakendada kvantitatiivset, matemaatilist mõtlemist keeleteaduses ja lähendada keeleteadust täppisteadustele" (Бодуэн де Куртенэ, 1963, 17). Nn. Moskva koolkonna keeleteadlased F. Fortunatov, M. Peterson jt. propageerisid statistiliste meetodite kasutamist vene keele grammatilise struktuuri uurimisel. Tartu Ülikoolis töötanud professor D. Kudrjavski teostas huvipakkuva statistilise uurimuse vene keele ajaloolise grammatika alal (vt. Кодухов, 1974, 246). On tuntud veel rida vene teadlasi, kes kasutasid statistilisi meetodeid keelenähtuste uurimisel (näit. Петров, 1911; Морозов, 1915). Eriti võib esile tõsta A. Peškovski töid vene keele häälikute statistilise struktuuri kohta (Пешковский, 1925). Metoodiliselt suur samm edasi oli nõukogude keeleteadlaste V. Tšistjakovi ja B. Kramarenko ühine uurimus statistiliste meetodite kasutamisvõimalustest keelematerjali põhjal (Чистяков, Крамаренко, 1929).

Sajandi esimesel poolel välismaal ilmunud keelestatistilistest töödest võib esile tõsta G. Dewey' (1923) põhjalikku uurimust inglise keele häälikusüsteemi kohta, E. Thorndike'i sagedussõnastikke (1921 jj.), rida pedagoogilise suunitlusega töid (keeleõpetuse ja õigekirjutuse alalt) ning P. Menzerathi (1944 jj.) fonotaktilisi uurimusi. Praha koolkonna esindajad V. Mathesius, B. Trnka ja N. Trubetzkoy töötasid välja statistilise uurimise põhimõtted fonoloogia valdkonnas.

Kõige olulisemad saavutused sel perioodil kuuluvad siiski matemaatikutele, kes püüdsid keelenähtuste statistilise analüüsi teel avastada üldisi seaduspärasusi teksti genereerimisel.

Juba 1913. aastal ilmus vene matemaatiku A. Markovi kuulus töö vene keele vokaalide ja konsonantide kõrvutiesinemise kohta Puškini teoses "Jevgeni Onegin" (Марков, 1913). Uurimus näitas, et teatud tingimuste korral võib küllaldase täpsusega ennustada kõrvutiesinemise tegelikke vorme. A. Markovi töö sai aluseks uuele matemaatilisele ajajärgule, mis põhineb tõenäosusteooria väljatöötamisel. Uurimus tõestas ühtlasi tõenäosuslike meetodite kasutamise võimalikkust keeleprobleemide lahendamisel. A. Markov oli ka esimene, kes osutas vajadusele toetuda keelenähtuste statistilisel uurimisel matemaatilise statistika ja tõenäosusteooria printsiipidele.

Tähtsa panuse keele statistilisse uurimisse tegi ameerika statistik G.K. Zipf, kes avastas mitmed olulised seaduspärasused teksti statistilises struktuuris. Kõige tuntum on nn. Zipfi seadus sõnasageduse ja sagedusjärgu seose kohta. Zipf uuris ka seoseid ja sõltuvusi sõnasageduse ja polüseemia vahel, häälikute muutumise alal jne., kusjuures teda huvitasid psühhofüsioloogilised faktorid, mis määravad inimese kõnetegevust. Zipfi arvates on kõnetegevuses olulise tähtsusega väljendusvahendite ökonoomia ("minimaalse jõupingutuse") printsiip, mida saab näidata ja tõestada statistiliste meetoditega (Zipf, 1929 jj.). Zipfi populaarsus keeleteaduslikes ringkondades ei rajane mitte tema keelefilosoofilistel vaadetel, vaid tema poolt avastatud tegelikel seaduspärasustel, mis on tänapäeval saanud juba klassikalisteks. Zipfi avastuste mõjul toimus keelestatistika "ma-

tematiseerumine". Hakati vähehaaval opereerima selliste mõistetega nagu statistiline jaotus, jaotuse parameetrid, juhuslik suurus, tõenäosus jne. Ilmusid ka esimesed metoodilised tööd, mis tutvustasid keeleteadlasi statistilise analüüsi võtetega (näit. Reed, 1949).

Olulist osa keelestatistika arengus etendas tuntud inglise statistikateoreetik G.U. Yule, kes samuti nagu G.K. Zipf huvitus üldistest seaduspärasustest teksti kvantitatiivses struktuuris. G.U. Yule'i paelusid individuaalsed erinevused stiilides ja autorsuse kindlakstegemise küsimused. Siit lähtudes teostas ta hulgaliselt keelestatistilisi uurimusi, mis võimaldasid kindlaks teha mõningaid huvitavaid seaduspärasusi sõnavara statistilises struktuuris. Yule'i teeneks tuleb pidada ka seda, et ta formuleeris esimesena kõnetegevuse tõenäosuslikkuse kontseptsiooni, vaadeldes teksti kui statistilist kogumit (Yule, 1944).

1.2. Tänapäeva keelestatistika

Tänapäevase keelestatistika väljakujunemine on tihedalt seotud küberneetika rajamisega, informatsiooniteooria arenguga ning keeleteaduse uute rakendusvõimalustega. Päevakorda kerkisid mitmed praktilist lahendust nõudvad probleemid, nagu sidekanalite optimaalne kasutamine, automatiseeritud infootsisüsteemid, automaattõlkimine (masinatõlge) jm. Keelestatistilised uurimused osutusid nimetatud probleemide lahendamisel vajalikuks eeltööks ja kasulikuks abivahendiks. Peale selle suurenes keelestatistika osatähtsus ka keeleuurimistöös üldse, ilmusid uued põhjalikud uurimused foneetika (fonoloogia), morfoloogia ja süntaksi valdkonnas. Tulemusi kasutati keelte tüpoloogilisel uurimisel, funktsionaalsete ja individuaalsete stiilide võrdlemisel, võõrkeelte õpetamisel jne. Eriti arenes keelte ja allkeelte sagedussõnastike koostamine.

Juba viiekümnendail aastail hakati keele kvantitatiivsel uurimisel süstemaatiliselt rakendama matemaatilise statistika ja tõenäosusteooria meetodeid. Keelestatistilises uurimistöös kehtestati kindlad valiku- ja dosee-

rimispõhimõtted, kusjuures aluseks sai statistiline valimimeetod (väljavõttemeetod), mis võimaldab teha põhjendatud otsustusi andmete representatiivsuse ja usaldatavuse kohta. Eriti intensiivselt hakkas keelestatistika arenema kuuekümnendail aastail, mil asuti laialdaselt rakendama elektronarvuteid keelestatistilises uurimistöös ja andmete töötlemisel. Ilmusid mitmed tõsised uurimused keelestatistika teoreetiliste probleemide alalt (Guiraud, 1959; Herdan, 1960 jj.). Toimusid rahvusvahelised ja üleliidulised konverentsid, mis olid pühendatud keele kvantitatiivsele uurimisele (näit. Londonis 1952. a., Stokholmis 1969. a., Minskis 1969. a., Gorkis 1970.a., Kišinjovis 1971. a., Mahatškalas 1974. a.). Peaaegu kõigil keeleteaduse konverentsidel olid keelestatistika sektsioonid. Üleliidulise Teaduslike Ühingute Nõukogu juurde asutati 1972. a. keelestatistika komisjon. Paljudes maades töötavad praegu teaduslikud keskused ja laboratooriumid, mis spetsiaalselt tegelevad keelestatistika küsimuste lahendamisega. Ainuüksi Nõukogude Liidus on selliseid keskusi ja töörühmi kümmekond, keelestatistikarühmade juhtivatest jõudusest võib nimetada niisuguseid tuntud teadlasi nagu R. Piotrovski Leningradis, B. Golovin Gorkis, O. Sirotinina Saraatovis, N. Andrejev Leningradis, V. Perebeinos Kiievis, K. Bektajev Tšimkendis, T. Jakubaite Riias jt. Üks esimesi keelestatistika ja selle meetodite propageerijaid ning huvipakkuvate sõnavarastatistiliste uurimuste autoreid on R. Frumkina (Фрумкина, 1964), kes viimasel ajal on pühendunud psühholingvistiliste probleemide lahendamisele statistiliste meetodite kaasabil.

Nõukogude Liidus on viimastel aastatel kirjutatud ja kaitstud hulgaliselt keelestatistika-alaseid väitekirju, sealhulgas töid, mis käsitlevad statistiliste meetodite kasutamist keeleuurimise automatiseerimise ja masinatõlke valdkonnas (näit. Зубов, 1969; Джубанов, 1973). Nii Nõukogude Liidus kui välismaal ilmuvad mitmed keelestatistikale pühendatud kogumikud ja perioodilised väljaanded, nagu "Статистика речи" (Leningrad, 1968 jj.), "Статистичні параметри стилів" (Kiiev, 1967), "Statistical Methods in Linguistics" (Stokholm, 1961 jj.) jt. On ilmunud ka rida artikleid, õpikuid ja käsiraamatuid, milles käsitletakse statistiliste meetodite kasutamist keeleteaduses. Keelestatistiline literatuur on

paisunud niivõrd suureks, et on hakatud koostama bibliograafilisi kogumikke keelestatistiliste uurimuste kohta, sealhulgas annoteeritud väljaandeid (näit. Bailey, Doležel, 1968; Kvantitativní lingvistika, 1964 jj.). Juba esimene suurem bibliograafiateos (Guiraud, 1954) sisaldas 2500 nimetust. On ilmunud kaks bibliograafilist teatmikku keelestatistilistest töödest Nõukogude Liidus (Ермоленко, 1967; Бектаев, 1972). Võib nimetada ka mitmeid ülevaateid keelestatistika ajaloost ja eriti viimaste aastakümnete keelestatistikast (Harkin, 1957; Cohen, 1967).

Keeleuurimistöö automatiseerimine on tingitud vajadusest vähendada töömahtu ning -vaeva keeleüksuste loendamisel ja eriti andmete töötlemisel tänapäeva metoodika nõuete kohaselt. Tööde automatiseerimine on eriti aktuaalne sõnavarastatistika-alastes uurimustes ja sel alal on viimasel aastakümnel ka palju ära tehtud (vt. Mutt, 1966; Засорина, 1966; Josselson, 1967; vt. ka kogumikud " Автоматизация в лингвистике" Leningrad, 1966; "Les machines dans la linguistique", Prague, 1968). Perioodiliselt toimuvad rahvusvahelised konverentsid tekstide automaattöötluse küsimustes, näit. Grenoble'is 1967. a., Stokholmis 1969. a., kus tähtis koht on alati olnud ka statistiliste uuringute automatiseerimise probleemidel. Uus lingvistikaharu, mis tegeleb keele automaattöötluse küsimustega ja mida on hakatud nimetama i n f o l i n g v i s t i k a k s (Õim, 1974; vene keeles kasutatakse nimetust вычислительная лингвистика, инженерная лингвистика; inglise keeles - computational linguistics), vajab suurel määral keelestatistika abi. Küsimus seisneb selles, et keelelise informatsiooni automaattöötlusel ei saa toetuda ainuüksi matemaatilis-loogilistele meetoditele, vaid on vajalik "empiiriline, induktiivne lähenemine, kusjuures otsitakse kõige üldisemaid reegleid informatsiooniprobleemide lahendamiseks. --- Tuleb pöörduda ligikaudsete meetodite poole, mida kogemuste lisandudes saab täpsustada ja täiustada" (Maron, 1963, 144). Nii näiteks põhinevad peaaegu kõik sõnade automaatsegmenteerimise katsed statistilistel printsiipidel (Андреев, 1967; Иванова, Шайкевич, 1970; Пиквер, 1973). Ka masinatõlke probleemide lahendamisel on mitmed koolkonnad (näit. R. Piotrovski poolt juhitav "Kõnestatistika" uurimisgrupp) lähtunud tõenäosuslikest kriteeriumidest, mis saadak-

se keelestatistiliste uuringute tulemusena. Statistilisi meetodeid kasutatakse sõnaklasside automaatsel määramisel tekstides (näit. Перебейнос, 1971; Белоногов, 1974), nn. võtmesõnade kindlakstegemisel ja automaatindekseerimisel ning -refereerimisel (Carroll, Roeloffs, 1969; Lustig, 1969) ja paljude muude informaatika-alaste probleemide lahendamisel. Ei saa alahinnata nimetatud probleemide tähtsust teaduse praegusel arenguetapil, mil otsitakse uusi teid ja võimalusi teadusalase informatsiooni töötlussüsteemide parendamiseks.

Tehniliste rakenduste kõrval on keelestatistikal tähtis osa ka puhtlingvistilises uurimistöös. Keelestatistilisi andmeid kasutatakse väga paljudes uurimustes nii abistava näitematerjalina kui ka otseselt keele või kõne fragmentide mudelite loomiseks. Vaatleme lühidalt mõningaid tähtsamaid keelestatistika rakendusvõimalusi tänapäeva lingvistikas.

Keelte tüpoloogilisel ja kõrvutaval uurimisel ei piisa tavaliselt sellest, kui vaadeldakse foneetilisi nähtusi või sõnade tuletamist, muutumist, ühendamist lauseteks jne. kvalitatiivsel tasandil, vaid on vaja ka kvantitatiivseid karakteristikuid, et keeli tüpoloogiliselt süstematiseerida. Seepärast on loomulik, et statistilised meetodid leiavad laialdast kasutamist just keeletüpoloogilistes uurimustes ja on välja kujunenud uus lingvistika suund, mida võib nimetada k v a n t i t a t i i v s e k s k e e l e t ü p o l o o g i a k s . Nõukogude Liidus viljeleb seda suunda Leningradi teadlase N. Andrejevi poolt juhitav keelestatistikarühm, millesse kuulub teadlasi ka paljudest teistest linnadest. Välismaal tehtud töödest võib esile tõsta mitmeid uurimusi fonoloogilise ja morfoloogilise tüpoloogia valdkonnas (Menzerath, 1954; Greenberg, 1960; Kramský, 1966; Kučera, Monroe, 1968). Nii meil kui ka mujal otsitakse pidevalt uusi meetodeid ja võtteid kvantitatiivsete tüpoloogiliste uuringute tõhustamiseks (vt. näit. Mustonen, 1965, kus autor kasutab statistilist diskriminantanalüüsi keelte automaatseks eristamiseks statistiliste karakteristikute alusel). Viimaste aastate keelestatistiliste konverentside päevakorras on olnud üldistavaid ettekandeid, milles esitatakse uusi nõudeid keelte tüpoloogiliseks

ja kõrvutavaks uurimiseks statistiliste meetoditega (näit. Якубайтис, 1971).

Keelte tüpoloogilise uurimisega on lähedalt seotud allkeelte ja stiilide kvantitatiivne kõrvutav analüüs, mis kuulub k v a n t i t a t i i v s e s t i l i s t i k a (stilomeetria, stilostatistika) valdkonda. Funktsionaalsete ja individuaalsete stiilide uurimisel statistiliste meetoditega on silmapaistvaid tulemusi saavutanud nõukogude keeleteadlased B. Golovin, O. Sirotinina, V. Perebeinos, A. Šaikevitš, M. Kožina jt. (Vt. ka kogumikku "Вопросы статистической стилистики", Киев, 1974). Välismaistest uurijatest võib esile tõsta rea teadlasi Tšehhoslovakkias (J. Mistrík, M. Těšitelová, P. Vašák, J. Kraus ja ungari keelt uuriv T. Zsilka), Prantsusmaal (P. Guiraud, Ch. Muller), Saksa FV-s (D. Krallmann, H. Fischer, F. Antosch), Rootsis (S. Allén, H. Karlgren, J. Thavenius) jm.

Seoses vajadusega ratsionaliseerida k e e l t e õ p e t a m i s t on koostatud hulgaliselt grammatikaõpikuid ja miinimumsõnastikke statistiliste printsiipide alusel, kusjuures lähtutakse eri allkeeltest, mis on esmajärgulise tähtsusega antud keeleõppijale. Võetakse arvesse ka keelte erinevusi ning emakeele interferentsi. Sagedus- ja miinimumsõnastike koostamisel on suure töö ära teinud nõukogude uurijad, eriti üleliidulisse "Kõnestatistika" rühma kuuluvad teadlased R. Piotrovski, P. Aleksejev, L. German-Prozorova, V. Morozenko, I. Turuk jt. Võõrkeelte õpetamise optimiseerimise ja eri keelte sagedus- ning miinimumsõnastike koostamise alal on häid tulemusi saavutanud ka L. Hoffmanni poolt juhitav keelestatistikarühm Saksa DV-s Leipzigis (Hoffmann, 1969 jt.). Viimasel ajal on hakatud tähelepanu pöörama ka võõrkeelsete ning emakeelsete t e k s t i d e r a s k u s e (loetavuse, arusaadavuse, jõukohasuse) mõõtmisele, mida teostatakse nii psühholingvistiliste katsete kui ka teksti objektiivsete omaduste - statistiliste karakteristikute alusel (vt. Микк, 1974; Тулдава, 1975).

Statistilised meetodid leiavad laialdast kasutamist k e e l e k o n t a k t i d e uurimisel, sealhulgas laensõnade osatähtsuse vaatlemisel eri allkeeltes, bilingvismi probleemide lahendamisel jne. (Anttila, 1963; Stötzer, 1966; Смирнов, 1966; Соонтак, 1973). Hoogu on saanud ka statis-

tika kasutamine d i a l e k t o l o o g i a s ("murdestatistika"), kus uuritakse dialektide omavahelisi suhteid, algkodu küsimust, dialektide kujunemist ja muid probleeme (näit. Панкрац, 1968; Вейлерт, 1973; Мамсурова, 1974). Hästi tuntud on ungari teadlaste L. Papp'i ja V. Farkasi tööd murrete tüüpide määramisel statistiliste meetodite abil (Papp, 1963; Farkas, 1966). Murdestatistika seostatakse sageli lingvogeograafiliste uuringutega, nii näiteks on W. Doroszewski poola koolkond välja töötanud meetodi nn. kvantitatiivsete isoglosside uurimiseks (Ivić, 1969, 83).

Viimastel aastatel on paljud lingvistid jõudnud veendumusele, et statistika rakendamine võiks tuua suurt kasu d i a k r o o n i l i s e l e keeleuurimisele. Üks esimesi suundi sel alal oli ameerika lingvisti M. Swadeshi poolt rajatud "glotokronoloogia" (Swadesh, 1950). M. Swadesh töötas välja statistikal baseeruva metoodika, mis võimaldas kindlaks määrata nii keelte suguluse astet kui ka ligikaudset aega, mis on möödunud sellest, kui keeled lahknesid ühisest algkeelest. Glotokronoloogias kasutatava meetodi kohaselt vaadeldi nn. põhisõnavara muutumist eri keeltes aegade jooksul. Saadud tulemused äratasid algul suurt huvi keeleteadlaste ringkondades, kuid hiljem on avaldatud kahtlust glotokronoloogiliste uurimuste paikapidavuses. Tuleb aga märkida, et viimasel ajal on glotokronoloogia ideed kerkinud uuesti päevakorrale ning otsitakse uusi teid ja meetodeid vanade probleemide lahendamiseks (vt. näit. Lexicostatistics, 1973). Ka Nõukogude Liidus on viimasel ajal ilmunud mõned uurimused glotokronoloogia valdkonnast, neist võib nimetada M. Arapovi ja M. Hertzi tööd (Арапов, Герц, 1974) sõnavara muutumise seaduspärasuste kohta ja A. Piotrovskaja ning R. Piotrovski uurimust (1974) grammatiliste nähtuste ja sõnavara ajaloolise arengu alalt. Huvitav on märkida, et glotokronoloogia meetodit on rakendatud ka soome-ugri keelte ajaloolisel uurimisel (Raun, 1956; Fodor, 1960; Hajdu, 1962). Vastavad arvutused näitasid, et ungari keel ja läänemeresoome keeled lahknesid uurali algkeelest umbes neli ja pool tuhat aastat tagasi. Samuti on tehtud katset glotokronoloogia ehk "leksikostatistika" menetlust rakendada nn. altai teooria kontrollimisel (küsimus on selles, kas türgi, mongoli ja tunguusi-mandžu keeled pärinevad kõik

ühest altai algkeelest, vt. Клоусон, 1969).

Hulgaliselt on teostatud keeleajaloolisi uuringuid ka tavaliste keelestatistiliste meetoditega. Eriti huvipakkuvad on Saraatovi Ülikooli keelestatistikarühma tööd vene keele funktsionaalsete stiilide (ajalehekeele, teaduskeele jt.) arengu kohta käesoleva sajandi algusest tänapäevani (Сиротинина, 1968 jj.). Statistilisi meetodeid läti keele ajaloo uurimisel on kasutanud I. Freidenfelds (1967) ja A. Miķelsone (kandidaadiväitekirjas, 1967).

Viimase aja lingvistiliste uuringute seas võib nimetada ka tundmatute keelte ja vanade tekstide d e š i f r e e r i m i s t statistiliste meetodite kaasabil. On tuntud näiteks M. Ventrise ja J. Chadwicki katse dešifreerida kreeta silpkirja (lineaarkirja B), nõukogude teadlase J. Knorozovi maajade kirja dešifreerimine nn. positsioonilise statistika meetodi abil, M. Arapovi, A. Karapetjantsi jt. kõrvutavad statistilised uurimused nn. kidani tekstide mõistmiseks. Vanade tekstide dešifreerimise üldised põhimõtted on formuleerinud J. Knorozov ja M. Probst (1969).

Statistilistel meetoditel on tähtis koht p s ü h h o l i n g v i s t i l i s e s uurimistöös. Võib nimetada R. Frunkina koolkonna huvitavaid töid subjektiivsete sõnasagedushinnangute uurimisel (Фрумкина, Василевич, Герганов, 1971), A. Leontjevi ja G. Štšuri uurimusi sõnaassotsiatsioonide valdkonnas (Леонтьев, 1969; Щур, 1974) ning rida pedagoogilise ja sotsioloogilise kallakuga keelestatistilisi töid, näit. lapsekeele sõnavara uurimise alalt (Захарова, 1967). Väga olulist osa etendavad statistilised meetodid k õ n e p a t o l o o g i a uurimisel (Howes, 1964; Holstein, 1965; Фрумкина, Василевич, Добрович, 1971). Ka s e m a n t i k a a l a s t e s töödes on hakatud laialdaselt rakendama statistilisi meetodeid. On ilmunud mitmed semantilised sagedussõnastikud, peale selle kasutatakse statistilist vaatlust sõna tähenduste struktuuri ja selle muutumise uurimisel (Whatmough, 1954; Клименко, 1970), sõnarühmade analüüsimisel ja kõrvutaval vaatlemisel (näit. värve tähistavate sõnarühmade uurimisel), teksti semantiliste seoste väljaselgitamisel (Скороходько, 1974). Eriti tuleb rõhutada nn. distributiiv-statistilise meetodi osatähtsust semantiliste väljade kindlakstegemisel ja te-

sauruste moodustamisel, mis omavad tähtsust nii lingvistika kui ka informaatika seisukohalt (vt. näit. Шайкевич, 1963; Бородин, Козокина, 1971; Петрина, 1974).

Keelestatistiliste probleemidega on seotud ka statistiliste meetodite kasutamine luulekeele ja värsi uurimisel (Põldmäe, 1969; Doležel, 1965).

Statistiliste meetodite tungimine keeleteadusse ja eri uurimissuuna väljakujunemine, mida võib nimetada statistiliseks lingvistikaks ehk keelestatistikaks (lingvostatistikaks), on saanud teoks. Tänapäeval võib vaevalt veel kohata keeleteadlasi, kes eitaksid statistiliste meetodite kasutamise võimalust ja kasulikkust keeleuurimistöös. On saanud selgeks, et kui kvantitatiivsed keeleuuringud annavad ebahuvitavaid või triviaalseid tulemusi, siis tähendab see vaid seda, et uurija seadis endale tunnetuslikust seisukohast ebahuvitava ülesande või ei tunne küllaldaselt kaasaegse keelestatistika uurimismeetodeid. Sel juhul ei saa süüdistada statistikat ega kvantitatiivset lähenemisviisi, nagu ei saa seda teha ka kvalitatiivsete meetodite puhul, kui uurija ei seisa oma ülesande kõrgusel. Kvantitatiivsete meetodite tähtsust keele uurimisel on tunnetanud väga paljud keeleteadlased, kes ise tegelevad peamiselt traditsioonilise, kvalitatiivse keeleuurimisega. Iseloomulikud on näiteks sellised sõnavõtud:

"Tõsiasi, et keeles on nähtusi, mida saab loendada, muudab matemaatiliste meetodite kasutamise keeleteaduses seaduspäraseks. Matemaatilise aparaadi kasutamine õigustab end alati, kui see annab resultaate, mida teiste meetodite abil on raske või võimatu saada." (Филин, 1970).

"Kui vaadelda kvalitatiivse ja kvantitatiivse faktori osa keelearengus, siis on ilmne, et kvantitatiivset faktorit saab seostada keele funktsioneerimisega ja isegi sellega, mida nimetatakse keele ekstralingvistiliseks aspektiks. --- Sellepärast arvan ma, et keele funktsioneerimise uurimisel tuleb laialdaselt kasutada kvantitatiivseid, arvulisi meetodeid. See võimaldab seostada süsteemiväliseid ja süsteemisiseseid nähtusi, mis kogu keeleteaduse arengu jooksul on olnud teadlastele komistuskiviks." (Ярцева, 1964)

Paljud teadlased on rõhutanud statistiliste meetodite kasutamise vajadust eriti sellepärast, et "kvantitatiivsed

suhted iseloomustavad keelt oluliselt" (Строева, 1968) ja et "statistika täpsustab ja selgitab kvalitatiivseid probleeme, eriti neil juhtudel, kui tegelikkust ei saa otseselt kvalitatiivselt uurida, kas liigse keerukuse või heterogeensuse tõttu." (Trnka, 1949).

1.3. Keelestatistika Eestis

Esimesed teadaolevad statistilised andmed eesti keele kohta pärinevad A. Saarestelt (1932; 1952), kes teostas häälikute loendusi ja uuris eesti keele sõnavara etümoloogilist koosseisu kvantitatiivsest seisukohast. A. Saareste andmetel on eesti keele põhisõnavaras ümmarguselt 6000 sõnatüve ja neist umbes 60 % on soome-ugri algupära. Tavalises kõnekeeles moodustab soome-ugri osa isegi 80 %. A. Saareste andmed teksti kohta on aga saadud liiga väikese valimi põhjal (umbes 1000 sõnet, mille hulgas on 470 eri tüve, s. o. lekseemi), mistõttu uurimuse tulemusi võib pidada esialgseteks. Pealegi tuleb arvestada erinevusi eri allkeeltes. Teine suurem eesti keele alane uurimus vanemast perioodist on W. Andersoni sõnapikkuse statistiline vaatlus eesti rahvalaulus (Anderson, 1935).

Sõjajärgsel perioodil on Eestis statistilisi meetodeid rakendanud E. Laugaste (1969) ja A. Krikmann (1967) rahvalaulu uurimisel, J. Põldmäe eesti luule värsimõõdu uurimisel (1971), fonoloogia alal M. Hint (1969 jj.) ja K. Vende (1973); sõnavara alal S. Piir (1963), H. Vihma (1970), R. Reier (1969), H. Kasemets jt. (1970). Üks varasemaid keelestatistilisi töid grammatika valdkonnas oli E. Vääri uurimus verbi olema kasutamisest (Vääri, 1961). Statistilist vaatlust rakendab R. Kull eesti keele liitsõnade uurimisel (Kull, 1963). Vene keele kohta on silmapaistvaid keelestatistilisi uurimusi teostanud E. Šteinfeldt (TPedI) ja Z. Mints (TRÜ). Esimene neist on keeleõpetamise otstarbeks koostatud vene keele sagedussõnastiku autor (Штейнфельдт, 1963). TRÜ õppejõul Z. Mintsil on rida uurimusi poeetilise sõnavara alalt (Минц и др., 1967).

Huvitavaid tulemusi eesti keele morfoloogia statistilisel uurimisel on saavutanud H. Pak (Пак, 1965) ja H. Holmberg (Хольм, 1965), kes kuulusid N. Andrejevi poolt juhi-

tavasse keelestatistikarühma ja kasutasid oma uurimustes nn. "statistilis-kombinatoorset" meetodit (praegu nimetatakse seda meetodit "strukturaal-tõenäosusliku analüüsi" meetodiks).

Uut hoogu teoreetilisele ja praktilisele tööle tänapäeva keelestatistika nõuete kohaselt andis keelestatistikarühma asutamine Tartu Riiklikus Ülikoolis 1969. a. ja samal aastal korraldatud keelestatistika fakultatiivne erikursus ülikooli õppejõududele ja üliõpilastele, millest võttis osa ligi poolsada inimest. Esimesed uurimistulemused avaldati Ülikooli kogumikes "Linguistica" ja "Keel ja Struktuur" (alates 1969. a.). Esineti ettekandekoosolekutel ja kirjutati diplomi- ning võistlustöid eesti keele ja mõningate võõrkeelte statistilise uurimise teemadel. Autorid olid tookordsed TRÜ üliõpilased H. Niinemägi (1970), J. Valge (1970 jj.), P. Lääne (1969), I. Mullamaa (1970), T. Velliste (1971), L. Piller (1971), M. Linnamägi (1975) jt. Õppejõududest esinesid artiklitega S. Raitar (1972), J. Soontak (1970 jj.), N. Toots (1970 jj.), A. Pikver (1972 jj.), A. All (1972 jj.) ja mitmed teised. Neist J. Soontak, N. Toots ja A. Pikver kaitsesid väitekirja keelestatistika teemadel (võõrkeelte alal). Käesolevate ridade autor avaldas sarja artikleid statistiliste meetodite kasutamise kohta keeleteaduses ja teostas väiksemaid uurimusi eesti ja teiste keelte alal (Tuldava, 1969 jj.). Ülalnimetatud TRÜ keelestatistikarühma liikmete töödes on rarendatud matemaatilise statistika meetodeid ja võetud arvesse tänapäeva keelestatistikas kehtivaid nõudeid materjali valiku ja doseerimise, representatiivsuse ja statistilise usaldatavuse suhtes.

Keelestatistilisi uuringuid uute nõuete kohaselt hakati alates 1970. a. läbi viima ka Tallinna Pedagoogilises Instituudis dots. A. Villupi juhendamisel. Valmisid mitmed huvitavad uurimused eesti keele grammatika ja sõnavara valdkonnas (näit. Kaljund, 1970; Kesküla, 1972; Villup, 1972 jj.; Kõiva, Raadik, 1974; Tiits, Veiler, 1974; Enniko, Meiman, 1975). Huvipakkuvad on ENSV Pedagoogika Teadusliku Uurimise Instituudis teostatud statistilised uurimused kooliõpikute sõnavara kohta (Maanso, 1973 ja 1975).

Suurema ülesandena on TRÜ keelestatistikarühmal teoksil eesti keele sagedussõnastiku koostamine eri allkeelte järgi. See töö toimub käsikäes Tallinna Pedagoogilise Instituudiga (A. Villup) ja TRÜ arvutuskeskusega (Ü. Kaasik, K. Ääremaa). Suurt abi osutas töö algperioodil ka Tallinna Polütehnilise Instituudi arvutuskeskus (L. Võhandu, M. Räbovõitra) tekstide eeltöötlemise ja perforeerimise korraldamisel.

Väljaspool TRÜ filoloogiateaduskonna keelestatistikarühma on Ülikoolis viimaste aastate jooksul valminud paar uurimust eesti keele tähtede ja tähekombinatsioonide esinemissageduse kohta ilukirjanduse, ajaleht ja rahvaluulu tekstides (Kaasik, Laugaste, 1969, 1975). Nendes töödes kasutati elektronarvuti abi. Peale selle kavatsetakse TRÜ-s teha statistilisi uuringuid ka mõningate informaatika-alaste probleemide lahendamiseks (juriidilise kirjanduse erialakeele semantiliste ja temaatiliste sõnaväljade kindlakstegemine statistilis-distributiivse meetodi abil ning varem kvalitatiivsete meetoditega saadud tesauruse kontrollimine).

Lõpuks võib mainida, et keelestatistilisi uurimistöid eesti keele alal on tehtud ka välismaal. Ameerika Ühendriikides on uuritud eesti keele ühesilbiliste sõnade sagedusi sõnastikus (Raun, 1959) ning teostatud fonostatistilisi mõõtmisi (Lehiste, 1970). Rootsis koostati V. Tauli juhendamisel kirjanikusõnastik sagedusandmetega A. Mälgu romaani "Tee kaevule" põhjal (Tauli, 1964). Saksa FV-s on valminud "Õigekeelsuse sõnaraamatu" alusel eesti keele pöördsõnastik (Hinderling, 1975).

2. KEELESTATISTIKA TEOREETILISED ALUSED

2.1. Keelestatistika liigitus ja põhimõisted

Sissejuhatavas osas esitatud ülevaate põhjal selgusid keelestatistika rakendusvõimalused mitmesugustes eri valdkondades (infolingvistika, keeletüpoloogia, keeleõpetus, leksikograafia, stilistika, dialektoloogia, psühholingvis-

tika, kõnepatoloogia jt.). Uurimisobjekti järgi võib keelestatistikat jaotada mitmeks erinevaks alarühmaks. Fonostatistika uurib keele fonoloogilist süsteemi ja fonotaktilist struktuuri kvantitatiivsest seisukohast. Fonostatistikale on lähedane tähestiku- ehk grafeemostatistika, mis uurib tähtede sagedusi ja tähestiku statistilisi omadusi. Sõnade morfoloogilist struktuuri vaatleb morfeemostatistika, kusjuures eriline tähelepanu on pööratud sõnatuletusele ja sõna automaatsegmenteerimise probleemidele. Sõnade välist struktuuri vaatleb ka sõnapikkuse statistika. Leksikostatistika ehk sõnavarastatistika uurib sõnade esinemissagedust tekstis, sagedussõnastike omadusi jm.; võib vahet teha leksikoloogilise, leksikograafilise ja semantilise statistika vahel. Puhtgrammatiliste nähtuste kvantitatiivne uurimine kuulub morfoloogilise ja süntaksistatistika valdkonda. Süntaksistatistika alla kuulub ka viimasel ajal aktuaalseks saanud tekstilingvistika (fraasivälise lingvistika) probleemid, kuivõrd neid on võimalik kvantitatiivselt vaadelda. Nimetatud keelestatistilise uurimise aspektid võivad konkreetses vaatluses esineda ühendatult ja vastastikuses seoses, näit. keelte tüpoloogilisel uurimisel, stiilide analüüsimisel jne.

Nüüdisaegse keelestatistika põhiliseks meetodiks on valimimeetod (väljavõtteline vaatlus), mis põhineb matemaatilisel statistikal ja tõenäosusteoorial. Valimimeetodit kasutatakse neil juhtudel, kui tahetakse teha otsustusi terviku kohta selle terviku osa ehk nn. v a l i m i (väljavõtukogumi, väljavõtte)+ uurimise põhjal. Tervikut ennast nimetatakse ü l d k o g u m i k s (algkogumiks, populatsiooniks). Keelenähtuste uurimisel on valimimeetod kõige sobivam sel põhjusel, et keel on nn. lahtine süsteem, mida täpselt ei saa piirata ei kogu keele ulatuses ega ka üksikute allkeelte ulatuses, ning üldkogum pole tervikliкult uurimisele kättesaadav. Samuti on keeleuurimistöös tegemist suhteliselt suurte massiividega (keeleüksuste kogu-

+ Nimetus väljavõtukogum väljendab kõige täpsemalt vaatlusega hõlmatavat osa üldkogumist, kuid see termin on liialt pikk sageli esineva mõiste tähistamiseks. Termini valim kasutamisel ei tarvitse segada lähedus sõnale "valik" (sest ka näit "loodusliku valiku" puhul ei mõelda otseselt valimist). Valimi vasted võõrkeeltes on: vene k. выборка, saksa k. Stichprobe, inglise k. sample.

mitega), mis võimaldavad rakendada valimimeetodit ning samal ajal tagada valimi representatiivsust üldkogumi suhtes (lähemalt valimimeetodi põhimõtetest keelematerjali uurimisel vt. Tuldava, 1969).

Oluline küsimus keele statistilisel uurimisel on vaadeldava ü k s u s e täpne määratlemine. See peab toimuma igas konkreetses uurimuses vastavalt vaadeldavale materjalile. Võib konstateerida, et statistilise uurimuse seisukohast relevantsed keeleüksused moodustavad hierarhilise süsteemi eri tasandite näol, mida kujutatakse järgmiselt (vt. Андрющенко, 1969, 9 jj.):

foneetiline tasand ("null-tasand");
morfeemitasand (1. tasand);
sõnatasand (2. tasand);
süntagmaatiline tasand (3. tasand);
lausetasand (4. tasand).

Tasandite piirkondi võib täpsustada. Foneetilise tasandi alla kuuluvad tähed, häälikud ja foneemid ning nende ühendid, sealhulgas ka silbid. Morfeemitasandi moodustavad morfeemid. Sõnatasandi alla on koondatud sõnavormid ja lekseemid. Süntagmaatilise tasandi moodustavad sõnaühendid ja lausetasandi - laused.

Võttes aluseks ülaltoodud skeemi, võime defineerida teksti mõiste keelestatistika seisukohalt. T e k s t i all mõistame antud tasandi (i-nda tasandi) keeleüksuste jada, näiteks tähtede, foneemide, morfeemide, silpide, sõnade, lausete jada, olenevalt sellest, milliseid keeleüksusi me konkreetses töös otseselt vaatleme. Tekstis esinevate üksuste (tekstiüksuste) koguarvu antud tasandil saab vaadelda kui statistilist kogumit. Selle suurust nimetame teksti m a h u k s (ka teksti pikkuseks).

Loomuliku keele tekstis võivad keeleüksused reeglina korduda. See võimaldab meil moodustada tekstis esinenud eri keeleüksuste loendi, mis kujutab endast elementide "inventari". Sellist inventari nimetame vastavalt tasandile t ä h e s t i k u k s , s õ n a s t i k u k s vm. On võimalik loendada inventari elementide arvu ja kindlaks määrata inventari suurus ehk maht (näiteks sõnastiku maht). Järelikult on ka siin tegemist statistilise kogumiga (inventari-

üksuste kogumiga), mis aga erineb tekstiüksuste statistilisest kogumist selle poolest, et inventariüksused ei kordu.

Kui keelestatistilises uurimistöös on vaatluse all sõnatasand, siis mõistetakse t e k s t i all sõnaliste üksuste jada, kusjuures neid üksusi nimetatakse s õ n e d e k s (ehk tekstisõnadeks) ja nende arvu, s. o. teksti pikkust ehk mahtu tähistatakse tavaliselt tähega N. Formaalselt võib sõnet defineerida kui tähtede (häälikute, foneemide, silpide, morfeemide) järjendit kahe tühiku vahel. Sõned võivad olla kas kõik tekstis esinevad sõnad tavalises mõttes või - vastavalt uurimuse tingimustele - ühesilbilised sõnad, verbid, nimisõnad vm. Tekst võib seega koosneda ka ainult teatud liiki sõnaüksustest, mida "nopime" välja üldisest tekstist. Sõnatasandil mõeldakse s õ n a s t i k u all antud keeles (allkeeles, üksikus tekstis) esinevate eri sõnade loendit. Sõnastiku üksusi võib vaadelda kahest erinevast seisukohast: esiteks, arvestades tekstis esinevaid vorme (muutevorme) eri üksustena, näiteks venna, vennale, venda, mis annab meile s õ n a v o r m i d e loendi; teiseks, ühendades sõna muutevormid ühe nimetaja, tavaliselt põhivormi alla, näit. venna, vennale, venda → vend, mis annab s õ n a d e ehk l e k s e e m i d e loendi. Sõnavormide ühendamist lekseemi alla nimetatakse keelestatistikas "lemmatiseerimiseks" ja lekseeme vastavalt "lemmadeks". Sõnavormide arvu sõnastikus tähistame tähega V ja sõnade (lekseemide, lemmade) arvu tähega L. Neid termineid - sõne, sõnavorm, sõna (lekseem, lemma) - kasutame ainult siis, kui on vaja rõhutada vastavaid mõisteid. Teistel juhtudel piirdume tavalise üldnimetusega s õ n a .

Kui sõnavormidest või lekseemidest koosnevas loendis on antud ka vastavad esinemissagedused tekstis, siis kujutab selline loend endast s a g e d u s s õ n a s t i k k u .

Meie kogumikus hakatakse avaldama mitmesuguseid statistilisi uurimusi konkreetse keelematerjali põhjal. On loomulik, et eelnevalt tuleb peatuda üldistel teoreetilistel ja metodoloogilistel alustel, millel baseerub praktiline uurimistöö ja tulemuste analüüs ning interpretatsioon. Tuleb käsitleda statistiliste meetodite kasutamise põhjendatust keelelise materjali uurimisel, sõnastiku ja teksti vahekorda, meetodite valikut ja paljusid muid küsimusi, mille la-

hendamine on vajalik keelestatistilises uurimistöös. Nende küsimuste vaatlemine on eriti põhjendatud seetõttu, et seni on puudunud kokkuvõtlik keelestatistika teoreetiliste aluste käsitlus. Autor püüab alljärgnevalt süstematiseerida ja vajaduse korral täiendada olemasolevaid kontseptsioone ja neist järelduvaid praktilisi nõudeid keelestatistilise töö teostamisel.

2.2. Teooria osast keelestatistikas

Teatavasti on teooria küsimused keeleteaduses omandanud erilise aktuaalsuse alles seoses tänapäeva teaduse üldise arengusuunaga. Kauemat aega peeti teooriat keeleteaduses üleliigseks või vähetähtsaks ja pearõhk asetati empiirilisele vaatlusele ning meetodite väljatöötamisele ja katsetamisele. Samasugune olukord valitses ka keelestatistikas. Seepärast on loomulik, et keelestatistika pole jõudnud veel arendada ühtset ja terviklikku teooriat, kuid fragmentaarselt on mõnedki olulised üldistused ja põhimõtted juba avaldatud. Tähtsamaid neist püüame siinkohal kirjeldada ja süstematiseerida.

Teooria all tuleb mõista sellist loogilise mõtlemise vormi, mis väljendab kõige täielikumalt meie teadmisi mingi nähtuse kohta. Kuid teooria pole ainuüksi teadmine iseenesest, vaid teadmine ja selle rakendamine üheskoos, s.o. teadmine kui tunnetusliku ja praktilise tegevuse vahend (Брандес, 1975, 12). Keelestatistika teooria aineks on küsimus tõenäosuslik-statistilise lähenemise adekvaatsusest keelenähtuste uurimisel ja kirjeldamisel. Teooria ülesandeks on sel juhul formuleerida nimetatud "adekvaatsus" teadmise vormis ja esitada see teadmine ühtlasi tegevusprintsiibi kujul, s. t. praktilise keelestatistilise uurimise printsiibi (või printsiipide) kujul.

Üheks tähtsamaks teooria omaduseks on mitmekesisuse taandamine ühtsusele. Meie uurimuses tähendab see seda, et keelestatistikat vaadeldakse kui l i n g v i s t i k a objekti. Olenemata keelestatistiliste uurimuste mitmekesisusest ja erinevatest eesmärkidest on neil uurimustel ühine lingvistiline alus ja igasugune interpretatsioon jääb lõpp-

kokkuvõttes ikkagi lingvistika raamidesse.

Teaduslik teooria kujutab endast "tõelise teadmise süsteemi", mis on tuletatud kindlatest loogilistest printsiipidest ehk nn. teoreetilistest abstraktsetest premissidest (Kopnin, 1969, 132). Iga teadusliku teooria kohta kehtib nõue, et selles oleksid eristatud kaks "osahulka": süsteemi lähteteesid (väiksem osahulk) ja kõik ülejäänud, lähteteesidest tuletatud järeldused. Millised on lähteteesid, mis võiksid olla aluseks keelestatistika teooriale, ja milliseid järeldusi võib neist teha? Vaatleme neid küsimusi alljärgnevalt.

2.3. Keelenähtuste tõenäosuslik-statistiline olemus

Statistiliste meetodite kasutamine mingi objekti uurimisel ei ole tingitud mitte ainult teadmise tõenäosuslikust iseloomust, vaid peamiselt sellest, et tunnetusobjekt ise oma liikumises ja arenemises ning vastastikuses seoses teiste objektidega allub tõenäosuslikele seaduspärasustele (Штофф, 1972, 131). Keelenähtuste uurimisel statistiliste meetoditega peab seega omaks võtma või vähemalt mitte tagasi lükkama hüpoteesi, et keeleüksuste valik kõneprotsessis allub tõenäosuslikele seadustele.

Kogu senise keeleteadusliku uurimistöö kogemused lubavad väita, et keelenähtustele on objektiivselt omased mitmesugused kvantitatiivsed tunnused. Varjatud kujul tunnustavad seda kõik uurijad, nimelt kui kasutatakse selliseid kvantitatiivseid mõisteid nagu "sagedane", "harva esinev", "hulgaliselt", "tavaliselt" jne. Kuna aga sellistel mõistetel on väga üldine tähendus, siis pole nad küllalt usaldatavad selleks, et neid võiks arvestada statistilise keeleteooria alusena. Olulisem on empiirilisel teel kindlaks tehtud fakt, et kuigi keeles on palju nn. juhuslikku, ilmneb selle korduval kasutamisel teatav seaduspärasus, nimelt üks või teine keeleline nähtus esineb kindla sagedusega. On teada, et maailmas, milles me elame, valitsevad kahte laadi seadused - nn. dünaamilised ja statistilised (tõenäosuslikud). Esimest tüüpi seaduste toimet saab täpselt ette öelda, kuna aga teist tüüpi seaduste toimet võib ette ennusta-

da vaid teatava tõenäosusega, s. t. teatavates piirides, sest nende resultaadid kõiguvad pidevalt mingi keskmise suuruse ümber. Tõenäosuslikele seadustele alluvad oma arengus ja funktsioneerimises sellised looduslikud ja ühiskondliku elu nähtused, mis olenevad suurest hulgast erinevatest põhjustest, kusjuures need põhjused võivad olla erisuunalised või vastastikuses sõltuvuses ja seetõttu ei anna nad alati täpselt ühesugust resultaati. Kuid massilisel kordumisel lähenevad resultaadid mingile konstantsele suurusele, mida nimetatakse tõenäosuslikuks sageduseks ehk lihtsalt tõenäosuseks.

Ka keeleüksuste kasutamine kõneprotsessis sõltub tavaliselt nii suurest hulgast teguritest (lingvistilistest ja ekstralingvistilistest), et praktiliselt on võimatu neid kõiki arvestada. Seepärast saab keeleobjektide suhtes harva formuleerida täiesti determineeritud reeglit või seadust, kuigi mingi tendents on alati täheldatav.

Eelöeldu põhjal võib sõnastada faktorid, mis teevad võimalikuks lingvistiliste andmete statistilise vaatluse. Sellisteks faktoriteks on keeleliste lausungite massilisus, keeleobjektide korduvus nendes lausungites ja mingi kindla elemendi ilmumise juhuslikkus. Nimetatud faktorid (massilisus, korduvus, juhuslikkus) iseloomustavad tegelikult igasuguseid statistilisi süsteeme ja seepärast on loomulik, et analoogia põhjal teeme järelduse lingvistiliste kogumite statistilise loomuse kohta. Olukord on siiski keerulisem seetõttu, et kuigi võib tunnustada kahe esimese faktori - massilisuse ja korduvuse - paikapidavust lingvistiliste objektide suhtes, ei ole päris selge, mida tuleb mõista keeleelemendi ilmumise "juhuslikkuse" all. Küsimus seisneb selles, et keeles esinev juhuslikkus pole laadilt ühtne. Kui näiteks kõneleja-indiviid valib sõnu teatava konteksti tarvis, siis antud individuaalsel juhul on tegemist valikuga ja mitte juhuslikkusega. Kuid selline teadlik valik esineb koos juhuslikkusega. Esiteks tingib kõla (hääliku) ja tähenduse sõltumatus teineteisest seda, et valides sõnu tähenduse järgi, pole kõnelejal võimalik teostada valikut häälikute (foneemide) suhtes, mille esinemust määrab seega juhus.[+] Teiseks, sõnaesinemuste suur hulk (kusjuures korra-

[+] Siin arvestatakse juhusena ka nn. foneetilist sümbolismi, onomatopoeetilisi väljendeid jms.

takse palju vähemat hulka sõnavara üksusi, teeb võimalikuks vaadelda kõnes esinevat keeleelementide kogumit kui statistilist kogumit ja iga elemendi sagedust kui juhuslikku muutujat. Seega on sõnaesinemus määratud juhuse poolt. Ka paljude teiste ühiskondlike nähtuste uurimine on näidanud, et nn. tahtlikud aktid, kui neid vaadelda suurel arvul, alluvad statistilistele seadustele.

Ülalesitatud käsitlus juhuslikkuse osast kõneprotsessis vastab üldjoontes tuntud keelestatistiku G. Herdani (1956 jj.) kontseptsioonile keelest kui "valikust" ja "juhusest" (language as choice and chance). Herdan näeb valiku ja juhuse vastastikuses toimes "optimaalset süsteemi", millele läheneb ka loomuliku keele areng. Optimaalne süsteem on statistilist laadi selles mõttes, et see allub tõenäosusseadustele, mida modifitseerib süstemaatiline faktor. Selle põhjal võib öelda, et keel on juhus, kuid nii, et individuaalne kõneleja endale teadmata allub keele struktuursetele seadustele (Herdan, 1966, 11). Hääliku sõltumatus tähendusest on Herdani jaoks "aksioom nr. 1". Seepärast peab ka mittejuhuslik sõnade järjestus tekstis andma statistilises mõttes juhusliku valimi häälikutest, foneemidest, tähtedest. See kehtib Herdani arvates alati konkreetse keele ulatuses. Järelikult võib oodata häälikute, foneemide ja tähtede sageduste stabiilsust ühe keele piirides. Ka paljud grammatilised nähtused alluvad üldkeelelistele statistilistele seaduspärasustele. Eraldi tuleb aga käsitleda selliseid nähtusi nagu sõnade sagedus, lausepikkus jne., mis on suurelt osalt tingitud stiilist ja mille uurimiseks tuleb kasutada spetsiaalseid statistilisi meetodeid (Herdan, 1962, 23 jj.).

Hilisemad uurimused on näidanud, et diferentseeritud lähenemine eri keeletasandite statistilisele vaatlusele on kindlasti vajalik. Häälikute, foneemide ja tähtede statistiline "käitumine" erineb kahtlemata sõnade esinemusest kõnes ja vastavalt tuleb kasutada ka erinevaid meetodeid eri keeleobjektide statistilisel uurimisel. Väga oluline on siin statistilise üldkogumi mõiste. Statistilise käsitluse seisukohast peab mingisse üldkogumisse kuuluval nähtusel olema aprioorne tunnus - tõenäosus (tõenäosuslik sagedus), mis on stabiilne antud üldkogumi ulatuses. See tä-

hendab, et üldkogumit peetakse tõenäosuse suhtes homogeenseks. Keeleobjektide uurimisel eeldatakse, et vaatlustasandil kindlaks tehtud esinemissagedus vastab tõenäosuslikule sagedusele üldkogumis (kusjuures esinemissageduse hälve aprioorsest tõenäosusest ei ületa juhuslikkuse piire). Tegelikult aga määratakse üldkogum sageli kvalitatiivsest tunnusest lähtudes, näit. kogu keel, mingi allkeel, individuaalne keeletarvitus, isegi üksainus tekst. Mõningate keeleobjektide uurimisel võib aga selguda, et nende statistilised omadused ei vasta määratletud üldkogumi nõuetele. Keelestatistilise uurimise ülesandeks ongi sel juhul kindlaks teha, mil määral on õigustatud üldkogumi määratlemine antud keeleobjektide suhtes ja milliste kvantitatiivsete ja kvalitatiivsete meetoditega võib seletada olulisi hälbeid statistilisest ootuspärasusest.

Tuleks käsitleda veel küsimust sellest, kuidas teoreetiliselt modelleerida kõneprotsessi kui "juhuslikku protsessi".

Üks esimesi katseid luua teksti genereerimise teoreetiline mudel pärineb matemaatikult B. Mandelbrotilt (1957). Mandelbroti teooria kohaselt luuakse tekst tähtede ja sõnade ning nende vaheliste tühikute jadana, kusjuures keeleüksused valitakse juhuslikult, kuid erineva tõenäosusega. Lähtudes sellest oletusest näitas Mandelbrot, et sõnasageduste jaotumus vastab sel juhul valemile, mida tuntakse Zipfi-Mandelbroti seaduse nime all. Mandelbrotil oli ka teine teooria, milles ta lähtub analoogiast termodünaamikaga. Nii Mandelbroti teooriad kui ka mõned hilisemad kontseptsioonid (vt. lähemalt: Plath, 1961) annavad väga lihtsustatud kõneprotsessi mudeli. Mandelbrot ise tegi vahet "makrolingvistika" ja "mikrolingvistika" vahel. Esimene neist kujutab endast "suure-mastaabiliste" keelenähtuste statistilist uurimist. Makrolingvistika suhe mikrolingvistikasse (grammatikasse) on analoogiline termodünaamika suhtega üksikute gaasimolekulide mehaanikasse. Mõte on selles, et kuigi makroskoopiline kirjeldus iseenesest ei ole vastuolus mikroskoopilise kirjeldusega, ignoreerib see siiski mõningaid detaile molekulide käitumises alamal (mikroskoopilisel) tasandil. Kummati võimaldab makroskoopiline lähenemine termodünaamikas formuleerida mitmeid tähtsaid kvan-

titatiivseid seaduspärasusi, mida praktiliselt poleks võimalik olnud saada üksikute molekulide käitumist uurides. Analoogiliselt võib statistiline "makrolingvistika" saada kasulikuks instrumendiks suurte tekstimassiivide kirjeldamisel, mille puhul täielik ja detailne "grammatiline" töötlus oleks liiga raske ja keeruline. Kasutades tänapäeva terminoloogiat, võiks öelda, et makrolingvistika mudel lubab idealiseeritud kujul esile tuua mitmeid objekti omadusi, mis võimaldavad nähtuse olemust paremini tundma õppida.

Informatsiooniteooria looja C. Shannon (1951) vaatles teksti nn. ergoodilise Markovi protsessi realiseeringute kogumina. Siinjuures eeldatakse, et on olemas tõenäosus mingi märgi (tähe, silbi, sõna) ilmumiseks pärast gruppi, mis koosneb k märgist. Võib kõnelda sellest, et antud teksti genereerimine toimub sõltuvalt "eelajaloo" kuhjumisest. Loomulikult peab sel juhul möönma, et keeleüksuse ilmumine tekstis ei ole rangelt võttes sõltumatu sündmus. Tähtsamad statistilised jaotused, nagu normaaljaotus ja Poissoni jaotus, mida tavaliselt kasutatakse keelenähtuste uurimisel (vt. Вектаев, Лукьяненков, 1971), eeldavad aga sõltumatute juhuslike suuruste olemasolu. Et aga sõltuvate juhuslike sündmuste jaoks mõeldud matemaatilise aparaadi kasutamine keelestatistikas on seotud väga suurte raskustega (arvutuste keerukuse tõttu) siis lähtutakse keelestatistikas tavaliselt lihtsustatud eeldusest, et tekstis esinevad üksused on üksteisest sõltumatud. See viib meid tegelikult tagasi Mandelbroti mudeli juurde, kus eeldatakse, et modelleeritav tekst on statsionaarne juhuslik protsess (s. t. teksti genereeriv süsteem ja kogu tingimuste kompleks on muutumatud ajas) ja ei arvestata tõenäosuslikke seoseid elementide vahel. Selline seisukoht on õigustatud tõenäosusteooria ja matemaatilise statistika seisukohast. Nimelt on võimalik rakendada suurte arvude seadust sõltuvate juhuslike suuruste suhtes, kui nende omavahelise kauguse (lineaarses mõttes) suurenemisega sõltuvus nõrgeneb (vt. Вектаев, Лукьяненков, 1971, 62). Lingvistiline reaalsus vastab sellele nõudele. Kuigi sõnade ilmumine tekstis oleneb teatud määral stiilist, temaatikast ja muudest ekstralingvistilistest faktoritest, on tõenäosuslikud seosed üksiku-

te sõnade vahel sellist laadi, et need järjest nõrgenevad sõnade omavahelise kauguse (vahemaa) suurenedes. Eksperimentaalsed andmed räägivad sellest, et informatsiooniteoreetiliste meetoditega mõõdetud seosed tegelikult vaibuvad juba nelja-viie sõna järel (Пиотровский, 1968). Seega on matemaatiliselt õigustatud vaadelda sõnade esinemist tekstis juhuslike sõltumatute sündmuste jadana. Järelikult on õigustatud ka vastavate matemaatiliste (statistiliste) meetodite kasutamine keeleüksuste sageduste uurimisel tingimusel, et arvestatakse nõudeid valimi mahu, suuruste hajuvuse jne. suhtes.

Teksti genereerimise tõenäosuslik-statistiline mudel kujutab endast reaalse teksti lihtsustust, kuid lihtsustatud eelduste vastuvõtmine lubab meil kasutada statistilisi meetodeid mõningate oluliste probleemide lahendamisel. Senise keelestatistilise uurimistöö kogemused näitavad, et järeldused, mis on tehtud mudeli alusel, on paljudel juhtudel täiesti vastuvõetavad ka reaalse teksti suhtes. Loomulikult jääb püsima nõue, et tõenäosuslik-statistilist mudelit kontrollitakse iga kord konkreetse keelelise materjali põhjal.

Esitatud tõenäosuslik mudel ei ole ainuke võimalus kõneprotsessi matemaatilisel modelleerimisel. Olenevalt töö eesmärgist ja iseloomust võib toetuda ka mitmesugustele teistele kontseptsioonidele (vt. näit. Леонтьев, 1974). Käesoleval juhul on aga tõenäosuslik-statistiline mudel loomulikuks eelduseks m a t e m a a t i l i s e s t a t i s t i k a meetodite kasutamisele keeleuurimistöös. Tõenäosuslik-statistilist teksti genereerimise kontseptsiooni võib pidada keelestatistika teooria üheks lähteteesiks, millega loogiliselt liituvad mõned teised olulised printsiibid. Üks põhilisi küsimusi on siin keele ja kõne eristamine.

2.4. Keele ja kõne statistiline interpretatsioon

Keelestatistika teooria aluseks on kommunikatiivne k õ n e t e g e v u s , mis on ühtlasi keelestatistika esmane uurimisobjekt. Kuid igasugust tegevust, sealhulgas ka kõnetegevust, saab tegelikult uurida vaid tegevuse resultaadi najal, s. o. mingi konkreetse "elementaarjuhtumi" najal,

milles "tegevusprotsess on objektiveeritud" (Брандес, 1975, 13-14). Selliseks konkreetseks elementaarjuhtumiks võib olla eri keeleüksuste kogum (näiteks sõnastik) või keeleüksustest koosnev tekst (kirjalik või suuline). Kui näiteks uurimuse peamiseks ülesandeks on vaadelda kvantitatiivselt keele sõnavara, siis vastava allteooria (sõnavarastatistika teooria) objektiks tuleb pidada sõnastikku ja teksti. Sõnastik ja tekst pole aga teooria objektiks otseselt, vaid kaudselt, sest teoorial on teatavasti kaks uurimisobjekti - ideaalne ja reaalne. Otseseks teooria objektiks tuleb pidada ideaalset, mis kujutab endast mõttelist originaalijäljendit. Sõnastiku ja teksti "originaaliks" on sel juhul k e e l ja k õ n e . Keele ja kõne eristamine on tänapäeva keeleteaduses üks olulisemaid põhimõtteid, kusjuures aga keele ja kõne mõiste tõlgitsemine võib olla erinev vastavalt keeleteaduse suunale ja uurimiseesmärkidele (ülevaadet erinevatest tõlgitsustest vt. Rätsep, 1963; Звегинцев, 1973). Ka keelestatistikas on esinenud ja esineb erinevaid seisukohti keele ja kõne määratlemisel. Kuna see probleem on tihedalt seotud eespool esitatud kõneprotsessi tõenäosusliku mudeliga ja täiendab ning süvendab seda mudelit oluliselt, siis anname järgnevalt ülevaate põhilistest seisukohtadest keele ja kõne statistilisel interpreteerimisel.

Kõige üldisemalt võib küsimuse asetada nii, et kõnet samastatakse tekstiga, mis tõenäosusliku mudeli järgi on "juhuslik protsess", kusjuures teksti (kõne) üksusi vaadeldakse kui "juhuslikke sündmusi". Juhuslik protsess eeldab aga mingit g e n e r e e r i v a t s ü s t e e m i, mida nimetamegi keeleks. Seejuures võib vaadelda nii teksti kui keelt eri tasanditel (foneetilisel, morfoloogilisel jne.). Oluliseks tingimuseks on nõue, et keeles endas peavad olema antud keeleüksuste tõenäosused, mis on alati lähedased keeleüksuste suhtelistele sagedustele reaalsetes tekstides. Sellise käsitluse korral saab keelt vaadelda kui ü l d k o g u m i t ja kõnet kui v a l i m i t (väljavõtukogumit) vastavast üldkogumist. Nii mõistsid keele ja kõne vahekorda keelestatistika teoreetikud G. Herdan ja P. Guiraud. G. Herdan lähtus F. de Saussure'i keele ja kõne dihhotoomiast ja leidis, et katseliselt kindlakstehtud tekstisageduste stabiilsus eeldab kindlate esinemistõenäo-

suste olemasolu ka keele erinevatel tasanditel, s. t. keel (langue) ei sisalda mitte ainult üksuste inventari, vaid keeleüksusi koos tõenäosustunnustega (Herdan, 1956, 79). P. Guiraud' arvates ei saa keeleüksuse esinemissagedust tekstis vaadelda ainult kõne (parole) omadusena, vaid keele (langue) objektiivse tunnusena, millel on niisama suur tähtsus keele funktsioneerimise seisukohalt kui vormidel ja tähendustel. Keele ja kõne (teksti) vahekorda tuleb tõlgitseda nii, et "igasugune tekst kujutab endast mingi keeleseisundi peegeldust ja väljendab keele kvantitatiivset struktuuri ning semantiliste realisatsioonide võimalusi" (Guiraud, 1959, 17-18).

Huvitav on jälgida, kuidas G. Herdan käsitleb v a l i k u (choice) ja j u h u s e (chance) vahekorda keele ja kõne eristamisel. Tema arvates esineb juhus ainult kõnes, s. o. üldkogumist tehtud valimis. Üldkogumi, s. o. keele enda statistiline struktuur on aga määratud valikust, kuid mitte individuaalsest, vaid kollektiivsest valikust, mis on ajalooliselt kujunenud ühiskonna "lingvistilise aktiivsuse" tulemusena (Herdan, 1966, 28). Herdan mõistab seega keele olemust kui sotsiaalset nähtust, mis on inimühiskonna arengu produkt.

"Valiku" ja "juhuse" teesi on mõttekas seostada marksistliku filosoofia paratamatuse ja juhuslikkuse kategooriatega. "Paratamatus tuleneb nähtuste seesmisest olemusest ja tähistab nende seadust, korda, struktuuri" (Filosoofiline leksikon, 1965, 317). Juhuslikkus aga tekib paljude erinevate nähtuste koosmõju tulemusena. Täpsemalt öeldes, "iga nähtus tekib seesmise paratamatuse sunnil, kuid selle nähtuse tekkimine on seotud paljude välistingimustega, mis oma konkreetse omapära ja lõpmatu mitmekesisuse tõttu on juhuslikkuse, antud nähtuse juhuslike joonte ja külgede allikaks" (Samas, 318). Seepärast võib öelda, et seesmine paratamatus on alati seotud välise juhuslikkusega. See kehtib täiel määral ka keele ja kõne vahekorra kohta: keel kui väljakujunenud struktuur ja paratamatus realiseeritakse kõnes, mis allub juhuslikkuse mõjule ja millel on seepärast tõenäosuslik-statistiline struktuur. Et aga juhuslikkuse taga peitub alati paratamatus, siis on ka kõne põhimõtteliselt määratud keele poolt. Kõne, s. o. teksti oma-

duste uurimisel tuleb seda silmas pidada ja püüda juhuslikkusest tingitud mitmekesisuse taga avastada keele seaduspärasusi, sest "kus ... pealispinnal toimub juhuse mäng, seal valitsevad seda alati seesmised varjatud seadused, ja asi seisab ainult selles, et need seadused tuleb avastada" (K. Marx ja F. Engels, Valitud teosed, II kd., lk. 322).

Lähenedes keelestatistika teoreetilistele probleemidele dialektilise materialismi seisukohtadest, peame silmas pidama paratamatuse ja juhuslikkuse õiget vahekorda. See võimaldab meil keelenähtuste uurimisel arvestada nii staatilist kui statistilist ja takistab langemast äärmustesse (ühelt poolt ainult deterministlike seaduspärasuste tunnustamine ja teiselt poolt kõigi keelenähtuste seletamine ainult juhuslikkuse ja tõenäosuslikkusega).

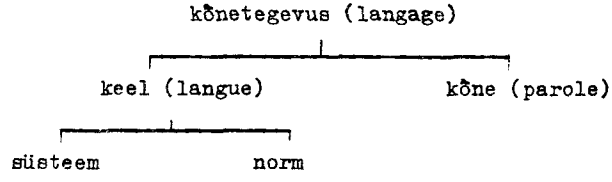
Nõukogude teadlaste R. Piotrovski ja L. Turõgina viimased uurimused on veelgi täpsustanud keele ja kõne mõistet ja funktsiooni keelestatistika valguses (Пиотровский, Турыгина, 1971). Autorid on seadnud endale ülesandeks välja selgitada, milline kahest tuntud skeemist - Saussure'i "keel - kõne" või Coseriu "keel - norm - kõne" - vastab paremini reaalsetele faktidele. Veenvalt läbiviidud eksperimendi tulemuste najal jõudsid uurijad järeldusele, et teksti statistilist struktuuri kujundab eriline e t a l o n ehk n o r m , mis asetseb mittestatistilise keelesüsteemi ja selle poolt genereeritava teksti (kõne) vahel. Kuna aga norm on Coseriu definitsiooni kohaselt keele (ja mitte kõne) komponent, siis pole Piotrovski ja Turõgina kontseptsioon tegelikult vastuolus ka Herdani vaatega, mille kohaselt tõenäosused kujutavad endast keele seesmist omadust. Herdani mudelit täpsustatakse lihtsalt uute elementide lisamisega.

Piotrovski ja Turõgina kontseptsiooni järgi võib keele ja kõne suhet lühidalt iseloomustada järgmiselt.

Keel on "informatsiooni edastamise koodsüsteem" (Пиотровский, 1970, 102). Teksti (kõne) genereerimisel piirab selle süsteemi võimalusi n o r m , mis kuulub süsteemi juurde ja kujutab endast kombinatoorset-tõenäosuslikku regulaatorit. Normi toime määravad nii psühhofüsioloogilised mehhanismid kui ka konkreetse keele omadused. Sellisel süsteemi ja normi k o o s m õ j u l genereeritaval tekstil on statistiline struktuur, mis hõlmab üldisi inimkeele iseära-

susi (näiteks keskmise sõnapikkuse diapasoon, informatsioonilise liiasuse aste) ja antud keele (allkeele, stiili, individuaalse stiili) spetsiifilisi omadusi.

Teatavasti arvestas F. de Saussure oma skeemis keele ja kõne kõrval ka k õ n e t e g e v u s t (langage). Kuid see skeem on omapärane selles mõttes, et loogilised seosed nimetatud kolme elemendi vahel pole ühesugused. Kõnetegevus jaotub kaheks vastandiks - keeleks ja kõneks, kusjuures ühelt poolt kõnetegevuse ja teiselt poolt keele ning kõne vahel on alluvussuhe, kuna aga keel ja kõne välistavad teineteist (dihhotoomia reegli põhjal). Kujutledes Saussure'i triaadi skemaatiliselt, peaksime seega triaadi liikmed asetama erinevatele hierarhilistele tasanditele. Lisades skeemile süsteemi ja normi komponendid (Coseriu mõttes ning Piotrovski ja Turõgina poolt keelestatistikale kohandatult), võime neid kujutleda alluvusvahekorras keelele ja vastastikuses välistusvahekorras. Selliselt on küsimust vaadelnud V. Bogdanov, kes näitlikustab kirjeldatud vahekordi järgmise seemiga (Богданов, 1973, 18):



V. Bogdanovit on huvitanud ka küsimus, kuidas toimub eri liiki tõenäosuste+ jaotumus vaadeldava skeemi elementide vahel. Sellel küsimusel on printsipiaalne tähendus keelestatistika teooria seisukohast. Ei ole kahtlust selles, et kõnet iseloomustab kõneüksuste suhteline esinemissagedus tekstis (valimi- ehk tekstisagedus), mis on seni olnud praktiliste keelestatistiliste uurimuste peamiseks huviobjektiks. Keeles on aga olukord keerulisem. Keel hõlmab süsteemi ja normi. Normi kohta on Piotrovski ja Turõgina uurimuste põhjal teada, et siin on maksev "normeeriv" statistiline

+ Eristatakse klassikalist, statistilist, loogilist, aksiomaatilist ja geomeetrilist tõenäosust.

tõenäosus, mis reguleerib tekstisagedusi vastavalt keele, allkeele, žanri jt. ajalooliselt väljakujunenud omadustele. Millist tõenäosust võib aga omistada süsteemile? Piotrovski ja Turõgina järgi on süsteem mittestatistiline, kuid ainult selles mõttes, et süsteemil puudub statistiline tõenäosus, millele suhteline sagedus läheneb katsete arvu suurenemisega (vt. Вентцель, 1969, 30-31). Süsteemile võib aga omistada klassikalist tõenäosust, mis eeldab süsteemi kõigi elementide võrdvõimalikkust (Tiit, 1968, 19). Lingvistiliselt on see tõenäosus mõttekas ainult siis, kui on teada elementide arv, mida saab kasutada tüpoloogiliste võrdluste puhul (näit. fonoloogilisel tasandil, vt. Sigurd, 1963). Klassikalise tõenäosuse mõõte on tarvis ka informatsiooniteoreetiliste vaatluste läbiviimisel, nimelt kui arvutatakse nulljärgu entroopia liiasuse kindlakstegemiseks (lähemalt vt. Tuldava, 1970, 3?8). Eksisteerivad ka elementide rühmade tõenäosused süsteemis, näit. vokaalide ja konsonantide suhteline osa häälikute üldarvust, eri sõnaliikide osakaal keele sõnastikus jm. Nii üksikute elementide kui ka rühmade tõenäosusi keelesüsteemis nimetab N. Andrejev "paradigmaatilisteks" tõenäosusteks, kuna aga tekstis esinevaid suhtelisi sagedusi vaatleb ta kui "süntagmaatilisi" tõenäosusi (Андреев, 1967, 17). N. Andrejevile annab paradigmaatiliste ja süntagmaatiliste tõenäosuste võrdlemine (suhte arvutamine) olulisi andmeid keele ja kõne vahekorra selgitamiseks kvantitatiivsel pinnal ja ta kasutab neid andmeid keelte tüpoloogilisel võrdlemisel. V. Bogdanov soovitab aga nimetust "süntagmaatilised tõenäosused" kasutada ainult keelesüsteemi kuuluvate elementide seostuse kohta (Богданов, 1973, 18) ja tekstis, s.o. kõnes esinevaid suhteid nimetada tavalise terminiga "suhteline (valimi-)sagedus". See võimaldab veelgi täpsustada Bogdanovi skeemis toodud hierarhilisi suhteid, nimelt saab süsteemile allutada paradigmaatika ja süntagmaatika, mis on omavahel välistussuhtes:

Keele ja kõne vahekorra täpsustamisega selguvad statistiliste meetodite kasutamise võimalused ja piirangud erinevate keeletasandite kvantitatiivsel uurimisel. On selge, et kõnes (tekstis), millel on valdavalt tõenäosuslik-statistiline struktuur, saab rakendada matemaatilise statistika meetodeid tekstiüksuste sageduste ning nende hajuvuste mõõtmisel. Ka norm on määratud statistilise tõenäosuse poolt ning seega on võimalik teksti uurida normi kindlakstegemise seisukohast mitmesuguste matemaatilise statistika hüpoteeside kontrollimise meetodite abil. Olenevalt valimi mahust ja andmetest tekstiüksuste statistilise jaotumuse kohta tuleb teksti ja normi uurimisel kasutada sobivaid parameetrilisi või mitteparameetrilisi kriteeriume (teste). Selline uurimine on tihedalt seotud stiili kvantitatiivse analüüsiga. Et norm kuulub keele alla, siis tuleb tegelikult mõõnda, et teksti ja normi ühisvaatlus on ühtlasi kõne ja keele kõrvutav uurimine. Kõne (teksti) põhjal tehakse kindlaks seaduspärasused, mida määrab keel vahepealse lüli - normi kaudu. Kuid eespool vaadeldud skeemi kohaselt kuulub keele alla ka süsteem, millel iseenesest puudub statistiline tõenäosus. Süsteemis osalev paradigmaatiline alljaotus on kirjeldatav tõenäosusteooria klassikalise variandi abil, kuid süntagmaatilise alljaotuse üksuste ja klasside (ka paradigmaatilise tasandi klasside) interpreteerimiseks ei ole olemas spetsiaalset matemaatilist aparaati (vrd. Богданов, 1973, 19). Kõne võrdlemine süsteemiga (näit. teksti võrdlemine sõnastikuga) peab toimuma eriliste meetodite abil, mida on keelestatistikas viimasel ajal hulgaliselt välja töötatud. Näiteks kasutatakse spetsiaalseid suhete indekseid (sõnastiku mahu ja teksti mahu suhe: V/N ehk "mitmekesisuse indeks", funktsionaalse koormuse ja informatiivse koormuse indeksid jm.). Seega on võimalik kvantitatiivselt uurida seoseid kõne (teksti) ja seda genereeriva keelesüsteemi vahel, kuigi neid lahutab kvalitatiivne erinevus tunnuste variatiivsuse seisukohalt: kõnet iseloomustavad pidevalt kõikuvad suhtelised sagedused, kuna aga keelesüsteemile on omane jäik paradigmaatiline tõenäosus (kõigi elementide võrdvõimalikkus) või väheelastne süntagmaatiline tõenäosus (elementide kombinatsioonivõimalused). Erilist huvi pakuvad sellised seosed keele ja kõne eri tasandite vahel, mida saab matemaatiliselt väljendada korrelatiivse

või funktsionaalse sõltuvusena (näit. sõnastiku juurdekasvu sõltuvus teksti pikkusest, vt. Захарова,1967). Kõnetegevuse kompleksne uurimine kvantitatiivsete meetoditega, eri tasandite võrdlemine ja vastastikuste seoste ning sõltuvuste kindlakstegemine võimaldavad lahendada nii rakenduslikke probleeme lingvistikas, informaatikas, pedagoogikas jne. kui ka mõningaid olulisi küsimusi keeleteooria valdkonnas. Nii näiteks lubab keele ja kõne statistiline interpretatsioon selgemalt piiritleda neid mõisteid elementide variatiivsuse seisukohast ja süvendada arusaamist keele hierarhilisest struktuurist. Mitmeid tähtsaid keelelisi seaduspärasusi on võimalik avastada ja täpselt formuleerida ainult kvantitatiivsete meetodite vahendusel tingimusel, et meetodeid kasutataks vastavalt konkreetsele keelematerjalile.

2.5. Keel ja allkeeled

Nii nagu keelt saab uurida erinevatel struktuuritasanditel (foneetilisel, leksikaalsel jne.), nii on võimalik seda vaadelda ka erinevate allkeelte seisukohast. Iga arenenud ja arenev loomulik keel koosneb nimelt . a l l s ü s t e e m i d e s t , mida ühendavad teatud üldised omadused, kuid millel on ka rida spetsiifilisi seaduspärasusi ja iseärasusi. Sellised allkeeled on näiteks lokaalsed ja sotsiaalsed murded. Allkeelteks nimetatakse traditsiooniliselt ka funktsionaalseid stiile (teaduskeel, ilukirjandus, publitsistika jt.), mille eristamisel lähtutakse keele funktsioonist ja kasutamise eesmärgist (Виноградов, 1955, 73). Seejuures võib funktsionaalset stiili kui allkeelt mõista mitmejärgulisena, näiteks teaduskeelt võib vaadelda esimese järgu allkeelena, eri teadusharude oskuskeeli aga teise järgu allkeeltena jne. (Erelt jt., 1971, 367). Põhimõtteliselt võib teatava järgu allkeeleks nimetada isegi individuaalset kõnepruuki. See tähendab, et allkeel on sünonüümne keele allsüsteemi mõistega, mis eeldab invariantset südamikku ja spetsiifilisi tunnuseid.+

+ Mõned keeleteadlased teevad vahet keele allsüsteemi kui üldisema mõiste ja allkeele kui allsüsteemi allklassi vahel, kusjuures allkeele olulist eristustunnust nähakse selles, et on olemas teatav grupp inimesi, kes kasutab seda allkeelt loomuliku ja võib-olla ainsa suhtlemisvahendina (vt. Valge, 1972).

Keele statistilisel uurimisel on väga oluline arvestada erinevate allkeelte ehk -süsteemide olemasolu. Eelnevalt oli juttu sellest, et statistiliste meetodite kasutamise tingimuseks on materjali homogeensus. Seepärast peab silmas pidama, et keel kui allkeelte kogum tervikuna ei ole reeglina statistiliselt homogeenne kõigi struktuuritasandite seisukohast. Eriti kehtib see leksikaalse tasandi suhtes. Statistiline vaatlus saab olla täiesti adekvaatne vaid allkeelte, näit. funktsionaalsete stiilide või isegi ainult individuaalse stiili tasemel. Keelestatistiliste uurimuste tähtsaks eeltingimuseks peaks seega olema uurimisala piiramine saavutamaks materjali maksimaalset homogeensust. Tsiteeritakse sageli akadeemik A. Kolmogorovi sõnu: "Statistika keeleteaduses peab olema võimalikult liigendatud." Mõeldud on siin seda, et statistiline vaatlus tuleks läbi viia alati kitsa allkeele piirides. See tähendab, et mõnel juhul - olenevalt struktuuritasandist - ei saa rääkida statistilisest homogeensusest ja stabiilsetest sagedustest isegi funktsionaalse stiili ulatuses (Спепак, 1974, 100). Kas sel juhul peab üldse loobuma statistiliselt mittehomogeensete kogumite uurimisest? Siin võib olla kaks erinevat küsimuse lahendust. Esiteks võib keelestatistilisel uurimisel katseliselt teha kindlaks piirkonna (žanri, allžanri, individuaalse stiili, konkreetse teose või teose osa), kus antud keelenähtuste uurimisel saab vastavate meetoditega konstateerida stabiilseid sagedusi, ja jäädagi selle piirkonna raamidesse. Sel juhul õigustab statistiliste meetodite kasutamine nähtuse uurimisel end täiel määral, kuid sageli langeb ära võimalus allkeelt vaadelda nendes piirides, mida määravad kasutussfäär ning traditsiooniliselt väljakujunenud kvalitatiivsed seisukohad. Teisel juhul võib aga lähtuda keelestatistikas ja eriti kvantitatiivses stilistikas kujunenud tavast, mille järgi võetakse aluseks mingi laiema ulatusega allkeel, näit. funktsionaalne stiil, ning tehakse eelkõige kindlaks erinevate keeleüksuste sageduste kõikuvuspiirkond (variatsiooniulatus) antud allkeeles. Opereerides hulgaliste valimite ja osavalimitega ning nendest saadud keskmiste sagedustega (mis sel juhul alluvad ligikaudselt normaaljaotusele), võime arvutada üldkeskmise usalduspiirid ning neid lugeda vaadeldava struktuuritasandi "tsentrumiks" või "nor-

miks" antud allkeeles. Selle alusel saab kindlaks teha konkreetsed tekstid, mis asuvad tsentrumi (normi) piirides, ja ülejäänud tekstid, mis moodustavad perifeeria või mis, teisiti väljendudes, oluliselt erinevad allkeele keskmisest. Selline jaotus tsentrumiks ja perifeeriaks põhineb objektiivsetel alustel ja võimaldab võrrelda ning rühmitada tekste ja stiile.+

Sageli kerkib probleem, kas on võimalik teostada statistilisi uurimusi " k o g u k e e l e " kohta sellistel struktuuritasanditel, mis teadaolevalt pole homogeensed kogu keele ulatuses. Näiteks võib küsida, kas on mõtet koostada mingi konkreetse keele koondsagedussõnastikku. On ju teada, et kogu keele sõnavara koosneb paljudest heterogeensetest kihtidest, mis pealegi on pidevas arengus ja muutumises. Tänapäeva keelestatistikas valitseb seisukoht, et sagedussõnastikke tuleb esmajoones koostada allkeelte kohta, kuid ei välistata ka kokkuleppelise koondsagedussõnastiku koostamise võimalust. Selline koondsagedussõnastik peab hõlmama sünkroonilises lõikes kõige olulisemaid allkeeli kindlaksmääratud proportsioonides. Sama põhimõte kehtib ka teiste struktuuritasandite statistilisel uurimisel. Huvipakkuvad on seejuures andmed allkeelte ühise osa ja erinevate osade kohta.

2.6. Materjali representatiivsus ja usaldatavus

Varem juba nimetasime, et keelestatistilistes uurimustes tehakse järeldusi mingi kindlaksmääratud ü l d k o g u m i kohta, kusjuures lähtutakse tavaliselt vaadeldava terviku osa ehk valimi uurimisest. Sel juhul kehtib nõue, et valim oleks küllalt representatiivne, s. t. peegeldaks tervikut ehk üldkogumit nii, et valimi põhjal tehtud järeldusi võiks pidada õigeteks kogu terviku (üldkogumi) suhtes. Representatiivsuse määravad materjali valiku tingimused, valimi maht, lubatav viga ja statistiline kindlus (usaldusnivoo). Tuleb silmas pidada, et valimimeetodi rakendamisel ei saa otsustada midagi absoluutse lõplikkuse ja kindlusega, vaid

+ Tehniliselt võib uuritava keelenähtuse keskmisi sagedusi üksikvalimis võrrelda üldkeskmisega (vahe olulisuse kindlakstegemiseks) ka selleks spetsiaalselt ettenähtud statistilise menetluse abil, vt. näit. Rasch jt., 1973, 96 jj.).

ainult selle kindlusega, mille me ise ette määrame. Põhimõtteliselt kuulub otsustus representatiivsuse üle vastava teadusharu kompetentsi (Tiit, 1971, 76). See tähendab, et ka keeleteaduses jääb statistiliste meetodite kasutamine lõppkokkuvõttes siiski keeleteaduse enda raamidesse ja meetodeid kasutatakse vastavalt praktilises töös väljakujunenud ja end õigustanud tavadele. Teistes teadusharudes kehtivaid reegleid ja piiranguid keelestatistilistesse uurimustesse mehaaniliselt üle kanda oleks printsipiaalselt väär. Vaatleme ükshaaval representatiivsust ja usaldatavust puudutavaid põhimõtteid, lähtudes keelestatistika seisukohtadest.

Esimene küsimus puudutab v ä l j a v õ t u printsiipi. Teatavasti kasutatakse matemaatilises statistikas mitut eri liiki väljavõtte: juhuslik ehk juhuväljavõtt, tüüp-, mehaaniline, seeria- ja astmeline väljavõtt (vt. Mereste, 1975, 320 jj.; Морозенко, 1969, 46). Puhtstatistiliselt on väljavõtuprintsiip õige siis, kui kõigile üldkogumi liikmetele on tagatud ühesugune tõenäosus sattuda valimisse. See nõue on kõige paremini täidetud juhuväljavõtu korral, näiteks, kui kasutatakse keeleüksuste valikul nn. juhuslike arvude tabeleid (selliseid tabeleid leidub matemaatilise statistika käsiraamatuis, näit. Tiit, 1971a, 207 jj., Mereste, 1975, 474 jj.). Juhuväljavõtt vastab ka teoreetilistele eeldustele kõnenähtuste tõenäosuslik-statistilise loomuse kohta, mida vaatlesime eespool. Seejuures ei tule aga unustada keele mitmekihilisust, s. o. erinevate struktuuritasandite ja erinevate allkeelte olemasolu. Üldkogumi mõistet ei saa alati rakendada kogu keele kohta, vaid tuleb eelnevalt piiritleda allsüsteem (allkeel), mida vaatleme üldkogumina. Seega peab keelestatistilist uurimust alustama suunatud valikust, mille alusel materjal liigitatakse rühmadesse kas kvalitatiivsete või varasemate uurimuste põhjal saadud kvantitatiivsete põhimõtete järgi. Näiteks määratakse kindlaks tekstide kuuluvus allkeelte, žanride, autorite järgi või kombineeritakse valik ajaliste jm. kriteeriumidega. On selge, et esimesel etapil osutub selline materjali liigitamine ja piiramine vajalikuks ning nähtust tuleb uurida just kitsama allkeele piirides, et kindlaks teha allkeele tõepäraseid statistilisi parameetreid. Allkeele piirides võib uuri-

mismaterjali valikut teostada juhuväljavõtu või kombineeritud väljavõtu alusel. Sageli ilmneb, et kõige õigem on toetuda nii kvantitatiivsetele kui ka kvalitatiivsetele põhimõtetele materjali (tekstide) representatiivsuse määramisel. Küsimus seisneb selles, et eelnev kvalitatiivne analüüs peab selgitama konkreetsete tekstide tüüpilisust antud allkeele või žanri raamides, kusjuures tuleb arvestada kunstilisi, sotsioloogilisi jt. faktoreid. Nähtavasti peab arvesse võtma ka žanride proportsioone (leviku mõttes) teatavas allkeeles, näit. romaani, novelli, jutustuse, reisikirjelduse jt. osatähtsust teatud perioodi ilukirjanduses. Selline lähenemine on täiesti kooskõlas juba mainitud põhimõttega, et representatiivsuse küsimus kuulub konkreetse teadusharu kompetentsi.

Olles seega eelnevalt täpsustanud uuritava materjali ja ühtlasi üldkogumi piirid, võime asuda tegelikule statistilisele vaatlusele. Kerkib küsimus, kas nüüd saab rakendada täiesti juhuslikku väljavõttu. Näiteks sõnavara uurimisel peaksime juhuslike arvude tabeli abil välja noppima üksikud sõnad vaadeldavast tekstist. Spetsiaalsed uurimused on näidanud, et selline konsekventne juhuslik väljavõtt annab tegelikult vähe paremusi võrreldes materjali valikuga lõikude kaupa, kusjuures lõigud võetakse juhusliku või mehaanilise väljavõtu alusel (näit. teatud ühesuguste vahemaade järgi). Olenevalt vaadeldavatest keeleüksustest kõigub lõikude ("portsjonite") optimaalne suurus 100 ja 1000 üksuse vahel (vt. Андреев, 1967; Алексеев, 1968; Головин, 1971). Võib veel nimetada, et tüüp-, mehaanilise ja juhusliku väljavõtu kõrval on keelestatistilistes töödes teatud tingimustel võimalik kasutada ka seeriaväljavõttu (Морозенко, 1969, 45) ja astmelist väljavõttu (Алексеев, 1968, 62). Ülalkirjeldatud väljavõtupõhimõtted on omaks võetud suurema osa nõukogude keelestatistikute poolt ja samalaadilist metoodikat rakendatakse ka välismaiste autorite töödes (näit. Hoffmann, 1968; Königová, 1965; Těšitelová, 1970).

Olulise tähtsusega on v a l i m i m a h u küsimus. Mida suurem on valimi maht, seda usaldusväärsemad on tulemused, kuid peab silmas pidama ka otstarbekohasuse ja ökonoomia printsiipi. Matemaatilise statistika meetodid lubavad meil kindlaks teha nn. r e p r e s e n t a t i i v -

s u s v e a (esindusvea) antud usaldusnivool (vt. lähemalt: Tuldava, 1969). Representatiivsusviga väheneb aeglaselt võrreldes mahu suurendamisega, näit. kui valimit suurendada k korda, siis viga väheneb keskmiselt $\sqrt{k}$ korda. Suurendades valimi mahtu 100 korda, saaksime seega viga vähendada ainult 10 korda. Seepärast on otstarbekohane algul kindlaks määrata, milline viga meid rahuldab ja selle alusel arvutada piisav valimi maht. Keelestatistilistes töödes peetakse tavaliselt küllaldaseks, kui representatiivsusviga (suhteline viga) 95%-lisel usaldusnivool on 10 - 20 %, kusjuures sõnasageduste uurimisel on lubatud suhtelise vea suurus 30 % ja isegi rohkem (Бектаев, Пиотровский, 1974, 189). Piisava valimimahu saame arvutada vastavate matemaatilise statistika valemite abil (Tuldava, 1969, 23 jj.). Praktilise töö kogemused on näidanud, et vajalik (piisav) valimi maht oleneb struktuuritasandist, millel toimub statistiline vaatlus. P. Aleksejev on kindlaks teinud huvitava seose vaadeldava keeleüksuse pikkuse ja valimi mahu vahel, nimelt mida pikem on üksus, seda suurem peab olema valimi maht (Алексеев, 1969, 20). See tähendab, et tähtede sageduste uurimisel võib piirduda väiksema valimiga kui näiteks silpide või sõnade sageduste vaatlemisel. Valimi maht ja ühtlasi lubatud representatiivsusviga olenevad ka uurimuse eesmärgist. Autorite stiilide võrdlemisel on ukraina keelestatistikud seadnud järgmised ligikaudsed piirid: foneetilisel tasandil vähemalt 7000 - 8000 foneemi või tähte, sõnatasandil vähemalt 10 000 sõnet, lausepikkuse uurimisel 25 000 sõnet igast teosest (Переобейнос, 1967, 29). Nende nõuete puhul kehtivad aga mõned reservatsioonid. Eeldatakse, et vaadeldav materjal, millest tehakse valim, on küllaldasel määral homogeenne. Kui sageduste hajuvus osavalimite vahel on suur, siis tuleb vastavalt suurendada ka valimi mahtu. Vastupidi võib väita, et väga ühtlase jaotusega sageduste korral võib valimi mahtu isegi vähendada. Näiteks, kui ilukirjandusteose sõnavara statistilisel uurimisel on vaja teha keskmiselt 10 000-sõneline valim, siis piirdudes ainult autorikõne või tegelaskõnega, võime valimi mahtu oluliselt vähendada. Seejuures kehtib muidugi nõue, et valim koosneks väiksematest osavalimitest (lõikudest, "portsjonitest"), millest oli juttu eespool. Keeleüksuste

ühendamisel klassidesse (näit. sõnaliikide vaatlemisel) võib mahtu vähendada, võrreldes uurimustega elementaarüksuste tasandil.

Keelestatistilistes töödes hinnatakse suuruste hajuvust (viga) tavaliselt 95%-lisel u s a l d u s n i v o o l ning statistilisi hüpoteese kontrollitakse vastavalt 5%-lisel o l u l i s u s n i v o o l (usaldusnivoo ja olulisusnivoo kohta lähemalt vt. Tuldava, 1969, 18 jj. ja 1970, 133 jj.). Sellised usaldusnivoo ja olulisusnivoo piirid on aga ainult kokkuleppelised ja võivad tegelikult varieeruda olenevalt materjalist ja töö eesmärgist. Eriti puudutab see statistiliste hüpoteeside kontrollimist, s. t. neid juhte, kui tahetakse kindlaks teha suuruste vahe olulisust või valimirea homogeensust. Täiesti õigesti märgib tuntud statistika teoreetik C. Gini, et statistilise olulisuse määramisel ei tohi olla formaalset lähenemist (Gini, 1971, 63). Statistiline olulisusnivoo ehk "vea tõenäosus" sõltub vaadeldavate nähtuste iseloomust ja praktilise töö vajadustest. Seda mõtet rõhutatakse paljudes matemaatilise statistika ja tõenäosusteooria käsiraamatutes. Näiteks, kui tuhandest kahurimürsust plahvatavad 999 ja üks ei plahvata, siis võib "vea" tõenäosuseks lugeda 0,001 ja pidada seda praktiliselt ignoreeritavaks. Kui aga samasugune tõenäosus (0,001) kehtib langevarjude avanemise puhul, siis ei saa sellist ohtlikku praagi tõenäosust ignoreerida. Niisamuti võib arutleda keelenähtuste statistilisel uurimisel, nimelt kas mingi vea tõenäosus on sisulistel kaalutlustel vastuvõetav või mitte. Keelestatistika praktikas on kujunenud tavaks, et opereeritakse peamiselt kolme põhilise olulisusnivooga (α): 0,05 (5 %), 0,01 (1 %) ja 0,001 (0,1 %), kusjuures suuruste vahet hinnatakse vastavalt "tõenäoliselt oluliseks", "oluliseks" ja "väga oluliseks" (Tuldava, 1970, 135). Nagu juba öeldud, peab ka siin arvestama konkreetseid uurimistulemusi. Kui näiteks funktsionaalsete stiilide võrdlemisel arvestada, et viiel juhul sajast võime eksida ($\alpha = 0,05$), väites, et vahe on oluline (teades eelnevalt, et stiilide erinevus oleneb suurel määral ekstralingvistilistest asjaoludest), siis autorite võrdlemisel teatud žanri piirides on otstarbekam seada rangem piir erinevuse kindlakstegemisel, näiteks 0,01, s.t. võime eksida ainult ühel juhul sajast. Teisiti öeldes, kui olulisusnivoo ei ulatu

0,01-ni, võib väita, et võrreldavad suurused kuuluvad ühte ja samasse üldkogumisse. Mõned keelestatistikud seavad veelgi rangemad piirid keelenähtuste erinevuse hindamisel, näiteks G. Herdan soovitab kasutada olulisusnivood 0,003 (Herdan, 1966, 43). Kõige õigem on muidugi läheneda diferentseeritult vahe olulisuse ja homogeensuse hindamisele olenevalt stiili- ja struktuuritasandist, nagu seda on teinud näit. ukraina keelestatistikud (Перебейнос, 1967).

Tulles veelkord tagasi valimi mahu probleemi juurde, võime äsjaöeldut arvesse võttes väita, et ka selles küsimuses tuleb lähtuda materjali sisulisest analüüsist ja uurimuse eesmärgist. Väärivad tähelepanu uusimad katsed rakendada valimi mahu reguleerimisel A. Waldi iteratsioonimeetodit (vt. näit. Попеску, 1972). Meetodi eesmärk seisneb selles, et uurides materjali väikeste annuste kaupa ja igal etapil kontrollides tulemuste usaldatavust, võime oluliselt vähendada kogu valimi mahtu ja seega saavutada töö ja aja kokkuhoidu (valimi maht loetakse piisavaks, niipea kui pideval lähenemisel saavutatakse vajalik tulemuste usaldatavus).

Kokkuvõttes võib järeldada, et materjali representatiivsuse ja tulemuste usaldatavuse hindamisel peab lähtuma vaatlusmaterjali sisulisest analüüsist, uurimise eesmärgist ja statistilise uurimistöö tehnilistest nõuetest. Kasulik on lisaks arvestada veel nõuet, et uurimuse tulemused oleksid "taastekitatavad" (Фрумкина, 1971, 53), s. t. et tulemusi kontrollitaks uute eksperimentide abil.

Keelestatistilises uurimistöös omab suurt tähtsust ka küsimus nn. statistilistest jaotustest, millest oleneb rida praktilisi ja teoreetilisi seisukohavõtte uurimiste käigus. Statistilisi jaotusi keeleteaduses ja nende praktilist arvutamist vaatleb autor üksikasjalikult eri artiklis kogumikus "Linguistica", VIII (sarjas "Statistilised meetodid keeleteaduses").

Viidatud kirjandus

Anderson, W., Studien zur Wortsilbenstatistik der älteren estnischen Volkslieder. - Eesti Rahvaluule Arhiivi Toimetised, nr. 2. Tartu, 1935.

Anttila, R., Loanwords and Statistical Measures of Style in the Towneley Plays. - Statistical Methods in Linguistics, 2. Stockholm, Skriptor, 75-93.

Bailey, R.W., Doležel, L., An Annotated Bibliography of Statistical Stylistics. Ann Arbor, 1968.

Busemann, A., Die Sprache der Jugend als Ausdruck der Entwicklungsrhytmik. Jena, 1925.

Carroll, J.M., Roeloffs, R., Computer Selection of Keywords Using Word-Frequency Analysis. - "American Documentation". Vol. 20, No. 3, 1969, 227-233.

Cohen, M., Sur l'histoire de la statistique en linguistique. - Etudes de linguistique appliqué, No. 5. Paris, 1967, 3-8.

Dewey, G., Relative Frequency of English Speech Sounds. Harvard University Press, Cambridge, Mass., 1923.

Doležel, L., Zur statistischen Theorie der Dichtersprache. - Mathematik und Dichtung. München, 1965, 275-293.

Enniko, K., Meiman, S., Sõnaliikide kvantitatiivne esinemus A.H. Tammsaare romaani "Tõde ja õigus" I köite autorikõnes. TPed.I kursusetöö, juh. A. Villup. Tallinn, 1975.

Erelt, T., Kull, R., Põlma, V., Raiet, E., Torop, K., Keelekorraldus ja liitsõnad. - "Keel ja Kirjandus", 1971, nr. 6, 367-374.

Farkas, V., Fonémastatisztikai problémák a nyeljárástipustörténetben. - "Nyelvtudományi értekezések", 55, 1966.

F o d o r , J., A statisztikai módszer alkalmazásának néhány kérdése. - "Magyar nyelvör", 1960, Nr. 2.

F r e i d e n f e l d s , I., Prievārdu lietošanas biežums latviešu laikrakstos. - Leksikologijas un leksikogrāfijas jautājumi. Referātu tēzes. Rīgā, 1967.

F ö r s t e m a n n , E., Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen. - "Zeitschrift für vergleichende Sprachforschung, begr. von A. Kuhn". Bd. 1. Göttingen, 1852, 163-173.

G r e e n b e r g , J.H., A Quantitative Approach to the Morphological Typology of Language. - "International Journal of American Linguistics". Vol. 26, No. 3, 1960, 178-194.

G u i r a u d , P., Bibliographie critique de la statistique linguistique. Utrecht, 1954.

G u i r a u d , P., Problèmes et méthodes de la statistique linguistique. Dordrecht, 1959.

H a j d u , P., Finn-ugor népek és nyelvek. Budapest, 1962.

H a r k i n , D., The History of Word Counts. - "Babel". Vol. 3, No. 3. Bonn, 1957.

H e r d a n , G., Language as Choice and Chance. Groningen, 1956.

H e r d a n , G., The Calculus of Linguistic Observations. The Hague, Mouton, 1962.

H e r d a n , G., The Advanced Theory of Language as Choice and Chance. Berlin - Heidelberg - New York, 1966.

H i n d e r l i n g , R., Rückläufiges Estnisches Wörterbuch. I Das Material der Grundformen. Ragensburg, 1975 (mimeogr.).

H i n t , M., Phonostatistics Based upon Texts from Estonian Fiction. - Generatiivse Grammatika Grupi aastakoosoleku teesid. Tartu, TRÜ, 1969, 8-11.

H o f f m a n n , L., Zur Spezifik der Fachsprache in sprachstatistischer Sicht. - "Fremdsprachenunterricht", 1968, Nr. 11, 469-475.

H o f f m a n n , L., Die Bedeutung statistischer Untersuchungen für den Fremdsprachenunterricht. - "Glottodidactica". Vol. 3/4. Poznań, 1969, 47-81.

H o l s t e i n , A.P., A Statistical Analysis of Schizophrenic Language. - Statistical Methods in Linguistics, 4. Stockholm, Skriptor, 1965, 10-14.

H o w e s , D., Application of the Word-Frequency Concept to Aphasia. - CIBA Foundation Symposium on Disorders of Language. Ed. A. V. S. Reuck and M.O'Connor. London, 1964, 47-75.

I v i č , M., Keeleteaduse põhisuunad. Tartu, TRÜ, 1969.

J o s s e l s o n , H.H., Lexicography and the Computer. - To Honor Roman Jakobson. Essays on the Occasion of His Seventieth Birthday. The Hague - Paris, 1967, 1046-1057.

K a a s i k , Ü., L a u g a s t e , E., Tähtede sagedus eestikeelsetes tekstides. - "Keel ja Kirjandus", 1969, nr. 10, 600-605.

K a a s i k , Ü., L a u g a s t e , E., Ä ä r e m a a , K., Tähtede ja silpide sagedus eestikeelsetes tekstides. - "Keel ja Kirjandus", 1975, nr. 1, 21-29.

K a e d i n g , F., Häufigkeitswörterbuch der deutschen Sprache. Steglitz bei Berlin, 1898.

K a r l g r e n , H., Statistical Methods in Phonetics. - Manual of Phonetics. Edited by B. Malmberg. The Hague, 1968, 129-154.

K a l j u n d , H., ma- ja da-infinitiivide ja nende käändeliste vormide esinemissagedus A.H. Tammsaare romaanis "Tõde ja õigus" I. TPed.I. lõputöö (5.ptk.). Tallinn, 1970.

K a s e m e t s , H., T u i s k , V., V a h t r a m ä e , E., J. Liivi jutustuse "Vari" sõnavormide sagedussõnastik. TPed.I võistlustöö, juh. H. Vihma. Tallinn, 1970.

K e s k ü l a , E., Adverbi kui sõnaliigi kvantitatiivne esinemus eesti kaasaegse ilukirjandusliku proosa keeles. TPed.I lõputöö, juh. A. Villup, Tallinn, 1972.

K o p n i n , P.W., P o p o w i t s c h , M.W. (Hrsg.), Logik der wissenschaftlichen Forschung. Berlin, Akademie-Verlag, 1969.

K r á m s k ý , J., A Quantitative Analysis of Italian Mono-, Di- and Trisyllabic Words. - Travaux linguistiques de Prague, 1. Prague, Academia, 1966, 129-143.

K r i k m a n n , A., Keelestatistikat eesti vanasõnadest. - Emakeele Seltsi aastaraamat, 13. Tallinn, 1967, 127-153.

K u č e r a , H., M o n r o e , G.K., A Comparative Phonology of Russian, Czech and German. New York, 1968.

K u l l , R., Liitsõnade arenemiskulg viimase saja aasta jooksul. - Nonaginta. J.V. Veski 90. sünnipäevaks 27. juunil 1963. Emakeele Seltsi Toimetised nr. 6. Tallinn, 1963, 165-183.

Kvantitativní lingvistika. (Bibliography of Quantitative Linguistics.) Red. M. Těšitelová. Praha, 1964 jj.

K õ i v a , S., R a a d i k , E., Adverbi süntaktilised ja semantilised funktsioonid kaasaegse ilukirjandusproosa autorikõnes. TPed.I lõputöö, juh. A. Villup. Tallinn, 1974.

K ö n i g o v á , M., K otázce statistického výběru v lingvistice. - "Slovo a slovesnost", 26, 1965, No. 2, 161-168.

L a u g a s t e , E., Sõnaalguline ja sisealliteratsioon eesti rahvalauludes. Eesti rahvalaulu struktuur ja kujundid I. - TRÜ Toimetised, vihik 234. Tartu, 1969.

L e h i s t e , I., Temporal Organization of Spoken Language. - Working Papers in Linguistics, No. 4. Ohio State University, Columbus, Ohio, 1970, 95-114.

Lexicostatistics in Genetic Linguistics. Proceedings of the Yale Conference. Yale University, Apr. 3-4,

1971, Ed. Isidore Dyen. The Hague, Mouton, 1973. (Janua Linguarum. Series Maior, 69.)

L i n n a m ä g i , M., Versuch einer statistischen Analyse zweier Substile der deutschen Belletristik. - Linguistica, VI. Tartu, TRÜ, 1975, 61-75.

L u s t i g , G., The Devel_opment of an Automatic Indexing System at Euratom. - 1968 Meeting of European Library Workers on Nuclear Field. Brussels, 1969, 61-74.

L ä ä n e , P., Statistische Analyse einer Mikrosprache. - Linguistica, I. Tartu, TRÜ, 50-61.

M a a n s o , V., IV-V klassi õpikute leksika. Teaduslik aruanne P003617. Tallinn, 1973. (Käsikiri ENSV Pedagoogika Teadusliku Uurimise Instituudis.)

M a a n s o , V., Sõnavaraline töö nõuab tähelepanu. - Emakeeleõpetuse küsimusi, V. Tallinn, "Valgus", 1975, 83-102.

M a n d e l b r o t , B., Structure formelle des textes et communication. - "Word". Vol. 10, 1954, No. 1, 1-27.

M a n d e l b r o t , B., Linguistique statistique macroscopique. - Rmt.: Apostel, L., Mandelbrot, B., Morf, A., Logique, langage, et théorie de l'information. Paris, Presses Universitaires de France, 1957.

M a r o n , M., A Logician's View of Language-Data Processing. - Natural Language and the Computer. New York, 1963.

M e n d e n h a l l , T.C., The Characteristic Curves of Composition. - "Science", IX, No. 214, 1887.

M e n z e r a t h , P., Die Architektonik des deutschen Wortschatzes. - "Phonetische Studien". Heft 3. Bonn, 1954.

M e r e s t e , U., Statistika üldteooria. Tallinn, "Valgus", 1975.

M i ķ e l s o n e , A., Saikļu izlietojuma vēsturiskā attīstība latviešu valodā. Disertacija filol. kand. grāda iegūšanai. Rīgā, 1967.

M u l l a m a a , I., English Loan-Words in Swedish. TRÜ diplomitöö, juh. J. Tuldava. Tartu, 1970.

M u s t o n e n , S., Multiple Discriminant Analysis in Linguistic Problems. - Statistical Methods in Linguistics,4. Stockholm, Skriptor, 1965, 37-44.

M u t t , O., Masinaleksikograafiast. - "Keel ja Kirjandus", 1966, nr. 5, 295-301.

N i i n e m ä g i , H., Statistilise stiilianalüüsi probleeme. - Keel ja Struktuur, IV. Tartu, TRÜ, 1970, 136-141.

P a p p , L., Nyelvjárástörténet és nyelvi statisztika. Budapest, 1963.

P i i r , E., "Kalevipoja" sõnastik. - Teoses: Fr. R. Kreutzwald, Kalevipoeg. Tekstikriitiline väljaanne ühes kommentaaride ja muude lisadega, II. Tallinn, 1963, lisa IV, 246-402.

P i l l e r , L., Über die Häufigkeit der Wortarten im Deutschen. - Linguistica, III. Tartu, TRÜ, 1971, 179-189.

P l a t h , W., Mathematical Linguistics. - Trends in European and American Linguistics 1930-1960. Utrecht-Antwerp, 1961, 21-57.

P õ l d m ä e , J., Statistiline meetod nõukogude värsiteoorias. - "Keel ja Kirjandus", 1969, nr. 10, 591-599.

P õ l d m ä e , J., Eesti värsisüsteemid ja silbilis-rõhulise värsisüsteemi arengujooni XX sajandil. Väitekiri. Tartu, 1971.

R a i t a r , S., Über die syntaktische und semantische Information. - Linguistica, IV. Tartu, TRÜ, 1972, 134-142.

R a s c h , D., E n d e r l e i n , G., H e r r e n d ö r f e r , G., Biometrie. Berlin, VEB Deutscher Landwirtschaftsverlag, 1973.

R a u n , A., Über die sogenannte lexikostatistische Methode oder Glottochronologie und ihre Anwendung auf das Finnisch-Ugrische und Türkische. - Ural-Altaische Jahrbücher. Bd. 28, Heft 3-4. Wiesbaden, 1956.

R a u n , A., Monosyllabics in Estonian. - Ural-Altaische Jahrbücher. Bd. 31. Wiesbaden, 1959, 317-327.

R e e d , D.W., Statistical Approach to Quantitative Linguistic Analysis. - "Word". Vol. 5, 1949, No. 3, 235-247.

R e i e r, R., Aadu Hindi "Tuulise ranna" sõnavarast. TPed.I lõputöö, juh. A. Raielo. Tallinn, 1969.

R ä t s e p , H., Keele ja kõne eristamisest. - Nonaginta. J.V. Veski 90. sünnipäevaks 27. juunil 1963. Emakeele Seltsi Toimetised nr. 6. Tallinn, 1963, 243-255.

S a a r e s t e , A., Die estnische Sprache. Tartu, 1932.

S a a r e s t e , A., Kaunis emakeel. Vesteid eesti keele elust-olust. Lund, Eesti Kirjanike Kooperatiiv, 1952.

S h a n n o n , C.E., Prediction and Entropy of Printed English. - "Bell System Technical Journal". Vol. 30, 1951, 50-64.

S i g u r d , B., A Note on the Number of Phonemes. - Statistical Methods in Linguistics, 2. Stockholm, Skriptor, 1963, 94-99.

S o o n t a k , J., On the Role of Foreign Words in Swedish Sports Texts. - Linguistica, II. Tartu, TRÜ, 1970, 112-124.

S t ö t z e r , U., Zur Häufigkeit fremder Wörter in politischen Aufsätzen und Reden. - Wiss. Beiträge der Martin-Luther-Universität, Halle, 1966/7, F. 1.

S w a d e s h , M., Salish Internal Relationships. - "International Journal of American Linguistics." Vol. 16, 1950, 157-167.

T a u l i , V., Word Index to August Mälk's Tee kaevule I. The Institute of Finno-Ugric Languages. Uppsala, 1964.

T ě š i t e l o v á , M., On the Statistical Choice of Language Material for the Purpose of Lexical Analysis (from the Point of View of Random Sampling). - Prague Bulletin of Mathematical Linguistics, 1970, pt. 14, 39-60.

T h o r n d i k e , E. L., The Teacher's Word Book. New York, 1921.

T i i t , E., Tõenäosusteooria, I. Tartu, TRÜ, 1968.

T i i t , E., Matemaatiline statistika, I. Tartu, TRÜ, 1971.

T i i t , E., Matemaatilise statistika tabelid, I, Tartu, TRÜ, 1971a.

T i i t , E., Matemaatilise statistika tabelid, II, Tartu, 1972.

T i i t s , M., V e i l e r , L., Sõnade ja sõnaliikide sagedusest A.H. Tammsaare romaani "Tõde ja õigus" I köites (II ja III osasõnastik). TPed.I võistlustöö, juh. A. Villup. Tallinn, 1974.

T o o t s , N., On the Frequency of Occurence of the Stressed Vowel Phonemes in Present-Day English. - Linguistica, II. Tartu, TRÜ, 1970, 82-111.

T r n k a , B., K výstavbě fonologické statistiky. - "Slovo a slovesnost". 1965, No. 11, 59-64.

T u l d a v a , J., Statistiline väljavõttemeetod keeleteaduses. - Linguistica, I. Tartu, TRÜ, 1969, 5-49.

T u l d a v a , J., Informatsiooniteooria ja keeleteadus. - "Keel ja Kirjandus", 1970, nr. 6, 329-339.

V a l g e , J., Eesti keele käänete sagedused kolmes funktsionaalses stiilis. - Keel ja Struktuur, IV. Tartu, TRÜ, 1970, 145-161.

V a l g e , J., Ajalehekeele sõnavara statistiline analüüs. TRÜ diplomitöö, juh. H. Rätsep. Tartu, 1972.

V e l l i s t e , T., A Comparative Stylo-Statistical Study of Two Novels. TRÜ diplomitöö, juh. J. Tuldava, Tartu, 1971.

V e n d e , K., Phonetic Conditioning Factors of Pitch in Estonian Vowels. - Estonian Papers in Phonetics. Tallinn, 1973, 46-84.

V i h m a , H., Kirjanikusõnastik. - "Keel ja Kirjandus", 1970, nr. 11, 649-654.

V i l l u p , A., Adverbide esinemissagedusest. - Linguistica, IV. Tartu, TRÜ, 1972, 221-246.

V ä ä r i , E., Verbi olema statistikat. - "Keel ja Kirjandus", 1961, nr. 7, 409-411.

W h a t m o u g h , J., Statistics and Semantics. - Sprachgeschichte und Wortbedeutung. Festschrift A. Debrunner. Bern, 1954.

W h i t n e y , W.D., The Proportional Elements of English Utterance. - "Proceedings of the American Philological Association". Vol. 14, 1874.

Z i p f , G.K., Relative Frequency as a Determinant of Phonetic Change. - Harvard Studies in Classical Philology. No. 40. Cambridge, Mass., 1929.

Z i p f , G.K., The Psycho-Biology of Language. An Introduction to Dynamic Philology. Boston, Houghton Mifflen, 1935.

Y u l e , G.U., The Statistical Study of Literary Vocabulary. Cambridge University Press, 1944.

А л е к с е е в П.М. К вопросу о выборке в лингвистическом исследовании. - Частотные словари и автоматическая переработка лингвистических текстов. Тезисы докладов 2-ой межвузовской конференции. Минск, 1968, 8-11.

А л е к с е е в П.М. Некоторые вопросы теории и практики статистической лексикографии. - Статистика текста. Т. I. Минск, Изд. БГУ, 1969, 12-37.

А л л А. Предложное управление глаголов в немецком медицинском подъязыке. - Methodica, I. Tartu, TRÜ, 1972, 7-37.

Андреев Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языковедении. Л., "Наука",1967.

Андрющенко В.М. Статистическая структура текста как функция его представления. - Проблемы прикладной лингвистики. Тезисы межвузовской конференции. Ч. I. М., 1969, 9-15.

Арапов М.В., Херц М.М. Математические методы в исторической лингвистике. М., "Наука", 1974.

Бектаев К.Б. Статистика речи 1957-72 гг. (Библиографический указатель). Алма-Ата, 1972.

Бектаев К.Б., Лукьяненков К.Ф. О законах распределения единиц письменной речи. - Статистика речи и автоматический анализ текста. Л., "Наука", 1971, 47-112.

Бектаев К.Б., Пиотровский Р.Г. Математические методы в языкознании. Ч. 2. Математическая статистика и моделирование текста. Алма-Ата, 1974.

Белоногов Г.Г. Об использовании метода аналогии при автоматической обработке текстовой информации. - Проблемы кибернетики. Вып. 28. М., 1974, 239-244.

Богданов В.В. Статистические концепции языка и речи. - Статистика речи и автоматический анализ текста. Л., "Наука", 1973, 9-19.

Бодуэн де Куртене И.А. Избранные труды по общему языкознанию. Т. 2. М., 1963.

Бородин В.В., Козокина С.М. Построение графа совместной встречаемости слов на ЭВМ. - Вопросы лингвостатистики и автоматизации лингвистических работ. Вып. 5. Труды ЦНИИПИ, сер. 3. М., 1971, 59-67.

Брандес М.П. Информационно-регулятивная модель общей теории перевода. - Теория перевода и научные основы подготовки переводчиков. Материалы Всесоюзной научной конференции. Ч. I. М., 1975, 12-17.

Вейлерт А.А. Об использовании количественных данных в диалектологии. - "Вопросы языкознания", 1973, № 4, 119-123.

Вентцель Е.С. Теория вероятностей. Изд. 4-е, стереотипное. М., "Наука", 1969.

Виноградов В.В. Итоги обсуждения вопросов стилистики. - "Вопросы языкознания", 1955, № 1.

Головин Б.Н. Язык и статистика. М., "Просвещение", 1971.

Джубанов А.Х. Статистическое исследование казахского текста с применением ЭВМ (на материале романа М. Ауэзова "Абай жолы"). Автореф. канд. дисс. Алма-Ата, 1973.

Ермоленко Г.В. Тематическая библиография работ по лингвистической статистике на русском языке. Алма-Ата, 1967.

Засорина Л.Н. Автоматизация и статистика в лексикографии. Л., Изд. ЛГУ, 1966.

Захарова А.В. Опыт статистического исследования устной речи ребенка. - Исследования по языку и фольклору. Вып. 2. Новосибирск, 1967, 16-38.

Звегинцев В.А. Язык и лингвистическая теория. М., Изд. МГУ, 1973.

Зубов А.В. Переработка текста естественного языка в системе "человек - машина". Автореф. канд. дисс. Л., 1969.

Иванова Н.С., Шайкевич А.Я. Дистрибутивно-статистическое описание американских патентных текстов. - Вопросы лингвостатистики и автоматизации лингвостатистических работ. Труды ЦНИИПИ, сер. 3/70. Вып. 4. М., 1970.

Клименко А.П. Вопросы психолингвистического изучения семантики. Минск, "Вышэйшая школа", 1970.

К л о у с о н Дж. Лексикостатистическая оценка алтайской теории. - "Вопросы языкознания", 1965, № 5, 22-41.

К н о р о з о в Ю.В., П р о б с т М.А. Общая схема дешифровки исторических систем письма. - Проблемы прикладной лингвистики. Тезисы межвузовской конференции 16-19 декабря 1969 г. Ч. I. М., 1969, 154-155.

К о д у х о в В.И. Общее языкознание. М., "Высшая школа", 1974.

Л е о н т ь е в А.А. "Словарь стереотипных ассоциаций русского языка", его теоретические основы, задачи и значение для обучения русскому языку иностранцев. - "Вопросы учебной лексикографии". М., 1969.

Л е о н т ь е в А.А. Проблемы математического моделирования речевой деятельности. - В кн.: Основы теории речевой деятельности. М., "Наука", 1974, 73-80.

М а м с у р о в а Е.Н. Некоторые особенности каталанского языка во Франции в свете лингвогеографии и статистики. - "Вопросы языкознания", 1974, № 5, 117-123.

М а р к о в А.А. Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь. - Известия Импер. Академии Наук.Серия 6, т. 7. СПб, 1913.

М и к к Я. Методика разработки формул читабельности. - Советская педагогика и школа. Вып. 9. Тарту, Изд. ТГУ, 1974, 78-163.

М и н ц З.Г., А б о л д у е в а Л.А., Ш и ш к и н а О.А. Частотный словарь "Стихов о Прекрасной Даме" А.Блока и некоторые замечания о структуре цикла. - Труды по знаковым системам, 3. Уч. зап. ТГУ, вып. 198. Тарту, 1967, 209-316.

М о р о з е н к о В.В. О методе отбора текстов для статистического описания языка (на примере английской экономо-статистической литературы). - Статистика текста. Т. I. Минск, Изд. БГУ, 1969, 38-54.

М о р о з о в Н.А. Лингвистические спектры. Изд. отд. русского языка и словесности Академии наук. Т. 20, кн. I-4. I9I5.

П а к Х.Я. О некоторых статистико-комбинаторных характеристиках функциональных классов (на материале эстонского языка). - Статистико-комбинаторное моделирование языков. М.-Л., "Наука", I965, 483-489.

П а н к р а ц Г.Я. Нижненемецкий диалект в СССР. Докт.дисс. Л., I968.

П е р е б е й н о с В.И. отв. ред.: Статиситичні параметри стилів. Київ, "Наукова думка", I967.

П е р е б е й н о с В.И. Экспериментальное выделение семантических классов существительных с помощью электронной вычислительной машины. - Семантические проблемы автоматизации информационного поиска. Киев, "Наукова думка", I97I, 84-90.

П е т р и н а А.М. Алгоритм выявления группировок ассоциативных терминов, ранжированных по степени их значимости. - "Научно-техническая информация", сер. 2, I974, № 2, 22-27.

П е т р о в В. Звуковая характеристика французского языка по статистическим данным. - Уч. зап. Казанского ун-та. Казань. I9II.

П е ш к о в с к и й А.М. Десять тысяч звуков. - Методика родного языка, лингвистика, стилистика, поэтика. М., "Госиздат", I925.

П и к в е р А. Статистическая морфемная сегментация слов.- Linguistica, IV. Tartu, TRÜ, 1972, 122-133.

П и к в е р А. О применении дистрибутивно-статистического метода в морфемике (на материале английского языка). Автореф. канд. дисс. М., I973.

П и о т р о в с к а я А.А., П и о т р о в с к и й Р.Г. Математические модели диахронии и текстообразования. - Статистика речи и автоматический анализ текста.Л., "Наука", I974, 36I-400.

Пиотровский Р.Г. Информационные измерения языка. Л., "Наука", 1968.

Пиотровский Р.Г. Отраслевой вероятностный машинный перевод. – Статистика текста. Т. 2. Минск, Белорус. ун-т, 1970. 5–32.

Пиотровский Р.Г., Турыгина Л.А. Антиномия "язык – речь" и статистическая интерпретация нормы языка. – Статистика речи и автоматический анализ текста. Л., "Наука", 1971, 5–46.

Попеску А.Н. Последовательный анализ при автоматической аттрибуции текста. – Частные вопросы автоматического анализа текстов. Минск, 1972, 346–355.

Сиротинина О.Б. Некоторые жанрово-стилистические изменения советской публицистики. – Развитие функциональных стилей современного русского языка. М., "Наука", 1968.

Скороходько Э.Ф. Определение значимости элементов текста на основе сетевой модели. – Всесоюзный научно-технический симпозиум "Лингвистическое обеспечение автоматизированных систем управления и информационно-поисковых систем." Махачкала, 1974, 120–123.

Слепак Б.Я. О некоторых вопросах методики организации статистических исследований на синтаксическом уровне. – Структурная и математическая лингвистика, 2. Киев, "Вища школа", 1974, 99–105.

Смирнов И.Е. Статистика французских лексических заимствований в румынском языке. – Энтропия языка и статистика речи. Минск, 1966, 306–320.

Соонтак Я. Х. Английские заимствования в шведской прессе. Автореф. канд. дисс. М., 1973.

Строева Т.В. Сопоставительная статистика падежных форм имени существительного в немецком и русском языках. – "Иностранные языки в школе", 1968, № 5, 6–16.

Тоотс Н.Я. К проблеме о функциональной нагрузке ударных гласных фонем в диахронии английского языка. Канд. дисс. Тарту, 1972.

Тулдава Ю.А. Об измерении трудности текста. – Methodica. Труды по методике преподавания иностранных языков. Вып. 3. Уч. зап. ТГУ, вып. 345. Тарту, 1975, 102–120.

Филин Ф.П. Ленинизм и теоретические проблемы языкознания. – Доклад на юбилейной сесии общего собрания Отделения литературы и языка АН СССР 1 апреля 1970 г. Цитируется по реферату Ю.С. Елисеева "Вопросы языкознания", 1970, № 6, 133.

Фрумкина Р.М. Статистические методы изучения лексики. М., "Наука", 1964.

Фрумкина Р.М. Вероятность элементов текста и речевое поведение. М., "Наука", 1971.

Фрумкина Р.М., Василевич А.П., Герганов Е.Н. Субъективные оценки частот элементов текста как прогнозирующий фактор. – Вероятностное прогнозирование в речи. М., "Наука", 1971, 70–93.

Фрумкина Р.М., Василевич А.П., Добрович А.Б. Вероятностная организация речевого поведения в норме и патологии (при шизофрении). Опыт сравнительного исследования. – Вероятностное прогнозирование в речи. М., "Наука", 1971, 145–169.

Хольм Х.А. Выделение первого морфологического типа в эстонском языке на материале публицистических текстов. – Статистико-комбинаторное моделирование языков. М.–Л., "Наука", 1965, 219–224.

Чистяков В.Р., Крамаренко Б.К. Опыт приложения статистического метода к языкознанию. Вып. 1. Краснодар, 1929.

Шайкевич А.Я. Распределение слов в тексте и выделение семантических полей языка. – Иностранные языки в высшей школе. Вып. 2. М., 1963.

Штейнфельдт Э.А. Частотный словарь современного русского литературного языка. 2500 наиболее употребительных слов. Таллин, 1963.

Штофф В.А. Введение в методологию научного познания. Л., Изд. ЛГУ, 1972.

Щур Г.С. Теория поля в лингвистике. М., "Наука", 1974.

СТАТИСТИЧЕСКИЕ МЕТОДЫ И ЯЗЫКОЗНАНИЕ

Ю. А. Тулдава

Резюме

В статье излагается история развития лингвостатистики и рассматриваются возможности применения статистических методов в наши дни. Отдельная подглава посвящена лингвостатистическим исследованиям в Эстонии. Во второй части статьи рассматриваются теоретические основы лингвостатистики, дается обзор научных концепций по вопросам о статистико-вероятностной природе языка и речи, о статистической интерпретации некоторых важных лингвистических проблем (антиномия язык – речь, соотношение языка и подъязыков и др.) и делаются выводы для практического исследования языкового материала статистическими методами.

STATISTICAL METHODS AND LINGUISTICS

J. Tuldava

S u m m a r y

The article deals with the history of the development of linguostatistics as well as the possibilities for the application of statistical methods nowadays. A special section is devoted to linguostatistical studies in Estonia. The second part examines the theoretical foundations of linguostatistics, offers a survey of various scientific concepts of the statistical nature of language and speech, as well as of the statistical interpretation of some important linguistic problems (the antinomy of language and speech, relations between language and sublanguages, etc.). Some conclusions are drawn as regards the practical study of language by means of statistical methods.