● TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
● УЧЕНЫЕ ЗАПИСКИ ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
● ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS

VIHIK
377
ВЫПУСК

KEELE-STATISTIKA 1

TARTU
1976

TARTU RIIKLIKU ÜLIKOOLI TOIMETISED
УЧЕНЫЕ ЗАПИСКИ
ТАРТУСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS
ALUSTATUD 1893.a. VIHIK 377 ВЫПУСК ОСНОВАНЫ В 1893.г.

TÖID KEELESTATISTIKA ALALT

I

ТРУДЫ ПО ЛИНГВОСТАТИСТИКЕ

KEELESTATISTIKA

TARTU 1976

T o i m e t a j a i l t

Käesolevaga alustab ilmumist kogumik "Töid keelestatistika alalt", mis seab endale eesmärgiks statistiliste meetodite tutvustamise ja rakendamise keeleuurimistöös. Kogumikku toimetavad Tartu Riikliku Ülikooli Filoloogiateaduskonna keelestatistikarühma liikmed. Esimeses väljaandes avaldatakse teoreetilise sissejuhatava artikli kõrval materjale eesti keele ilukirjandusproosa statistilistest uurimustest. J. Tuldava ja A. Villupi artiklis käsitletakse eesti keele sõnaliikide statistikat. Nimetatud autorite ja Ü. Kaasiku ning K. Ääremaa ühistöönа valminud eesti keele ilukirjandusproosa autorikõne sagedussõnastikust avaldatakse I osa - sõnavormide sagedusloend.

О т р е д а к ц и о н н о й к о л л е г и и

Настоящий сборник является первым в серии работ по пропагандированию и применению статистических методов в языкознании. Сборник издается Группой лингвостатистики при филологическом факультете Тартуского государственного университета. В первом выпуске публикуются, кроме вступительной теоретической статьи, материалы статистического исследования эстонской художественной прозы. В статье Ю. Тулдавы и А. Виллуп дается статистический анализ частей речи эстонского языка. В сборнике публикуется первая часть частотного словаря авторской речи эстонской художественной прозы - частотный список словоформ (авторы: Ю. Каазик, Ю. Тулдава, А. Виллуп, К. Эяремаа). Написанные на эстонском языке статьи имеют резюме на русском языке.

Vom Redaktionskollegium

Der Sammelband "Beiträge zur Sprachstatistik" stellt sich die Aufgabe, die Leser mit den statistischen Methoden und ihrer Anwendung in der Sprachforschung vertraut zu machen. Die Ausgabe wird von den Mitgliedern der Gruppe der Sprachstatistik der philologischen Fakultät der Staatlichen Universität Tartu redigiert. Der erste Sammelband enthält neben dem einführenden Artikel auch Angaben über die statistischen Untersuchungen der schöngeistigen Prosa des Estnischen. So wird im Artikel von J. Tuldava und A. Villup die Statistik der Wortarten im Estnischen behandelt. Es wird auch der erste Teil des Häufigkeitswörterbuches der Autorenrede der estnischen schöngeistigen Prosa - die Häufigkeitsliste der Wortformen - als gemeinsame Untersuchung von obenerwähnten Autoren, Ü. Kaasik und K. Ääremaa veröffentlicht.

Editorial Note

The series "Linguostatistica" aims to publish articles on statistical methods and their application in linguistic studies. It is edited by the members of the Linguostatistics Group of the Faculty of Philology, Tartu State University. In addition to an introductory theoretical article the first issue contains materials on statistical studies of Estonian prose fiction. J. Tuldava's and A. Villup's article is devoted to the statistics of the Estonian parts of speech. Part 1 of a frequency dictionary of Estonian prose fiction - a frequency list of word forms - is likewise included. The dictionary has been compiled by the above mentioned authors in collaboration with Ü. Kaasik and K. Ääremaa.

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).

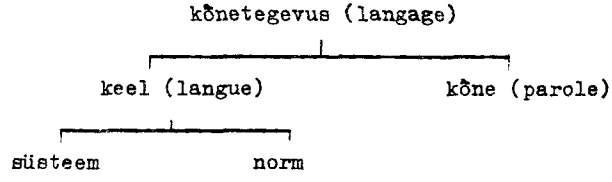
Nõukogude teadlaste R. Piotrovski ja L. Turõgina viimased uurimused on veelgi täpsustanud keele ja kõne mõistet ja funktsiooni keelestatistika valguses (Пиотровский, Турыгина, 1971). Autorid on seadnud endale ülesandeks välja selgitada, milline kahest tuntud skeemist - Saussure'i "keel - kõne" või Coseriu "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.

SÕNALIIKIDE SAGEDUSEST ILUKIRJANDUSPROOSA AUTORIKÕNES

J. Tuldava, A. Villup

1. SISSEJUHATUS

Sõnavara liigendamine leksikaal-grammatilisteks klassideks ja eri klasside funktsionaalse koormuse määramine on pidevalt püsinud keeleteadlaste huvi keskpunktis. Jättes kõrvale küsimuse sõnade klassifitseerimise eri võimalustest[+] ja võttes aluseks sõnaliikide traditsioonilise käsitluse (eesti keele kohta vt. Valgma, Remmel, 1968), vaatleme käesolevas uurimuses eesti keele sõnaliikide esinemust ühe allkeele - ilukirjandusproosa autorikõne piirides. Meetodina kasutame statistilist vaatlust ja analüüsi.

Sõnaliikide statistilise uurimise vajadust ja tähtsust on rõhutanud mitmed tuntud keeleteadlased. Juba 1938. a. kirjutas V.V. Vinogradov: "Sõnatüüpide esinemissagedused on ilmselt erinevad kirja- ja kõnekeele eri stiilides. Täpsed uurimused selles valdkonnas võimaldaksid kindlaks teha grammatilis-struktuurseid ja osalt ka semantilisi erinevusi stiilide vahel. - - - Grammatiliste kategooriate analüüs peab selgitama nende funktsionaalse kaalu eri stiilides." (Виноградов, 1938, 155-156). Hilisemad uurimused on veenvalt näidanud, et sõnaliikide sagedused ("funktsionaalne

[+] Sõnade liigitamise probleemi kohta lähemalt vt. Жирмунский, 1968; Стеблин-Каменский, 1974. Uusi liigitamise põhimõtteid esitavad näit. O. Sunik (Суник, 1966), I. Revzin (Ревзин, 1967), L. Sumarokova (Сумарокова, 1967, L. Andrejeva (Андреева, 1969). Strukturaal-lingvistikas on tuntud Ch. Fries'i (1952) distributiivsete sõnaklasside süsteem. Generatiivse semantika seisukohast on sõnaliike vaadelnud J. Lyons (1966), H. Õim (Ыйм, 1969) jt. Eesti keele sõnaklasside matemaatilis-lingvistilise käsitluse (hulgateooria alusel) esitab T. Tobias (Тобиас, 1962), statistilis-kombinatoorselt vaatleb eesti keele sõnaklasse H. Pak (Пак, 1965).

kaal") kujutavad endast tähtsaid kvantitatiivseid tunnuseid, mis võimaldavad eristada ja diagnoosida nii funktsionaalseid stiile ja žanre kui ka individuaalseid autoristiile (vt. näit. Клочкова, 1968; Тищенко, 1970; Кожина, 1972; Головин, 1974). Võrrelgem näiteks tähtsamate käändsõnaliikide esinemissagedusi tekstis vene keele eri funktsionaalsetes stiilides (Клочкова, 1968):

	Kõnekeel	Ilukirjandus (autorikõne)	Teaduslik kirjandus
Nimisõna	13,0 %	28,7 %	36,9 %
Omadussõna	3,5 %	7,9 %	16,4 %
Asesõna	17,5 %	10,2 %	6,2 %

Ka tegusõna osakaal varieerub tunduvalt vastavalt funktsionaalsele stiilile, näit. ukraina keele kohta tehtud uurimuse andmeil (Тищенко, 1970) on draamatekstis tegusõnu keskmiselt 22,0 %, ilukirjanduslikus proosas 19,7 %, teaduslik-tehnilises kirjanduses 13,9 % ja ühiskondlik-poliitilise sisuga tekstides 11,1 %. Analoogilisi andmeid sõnaliigisageduste varieeruvusest eri stiilides on saadud ka paljude teiste keelte materjali põhjal (vt. näit. Kelemen, 1964; Mistrík, 1969; Zsilka, 1973).

Eri funktsionaalsete stiilide (allkeelte, žanride) piirides on omakorda võimalik täheldada suuremat või väiksemat sõnaliigisageduste hajuvust keskväärtuse ümber, mis peegeldab individuaalseid erinevusi sõnaliikide kasutamisel. Üldreeglina on individuaalsed sagedusnäitajad määratud žanrilise kuuluvusega ja jäävad kindlatesse variatsioonipiiridesse. Vastavate meetoditega on aga võimalik kindlaks teha statistiliselt olulisi hälbeid "normist" või olulisi erinevusi autorite vahel. Individuaalsed erinevused on osaliselt tingitud välistest asjaoludest, näit. teema valikust-esituslaadist (jutustus, kirjeldus jne.), kuid suurel määral olenevad need ka autori isiksusest ja stiilimaneerist. Seetõttu on võimalik sõnaliikide sagedusi arvesse võtta autorsuse kindlakstegemisel (vt. näit. Тарлинская, 1969). Tuleb muidugi silmas pidada, et autori stiil võib aja jooksul muutuda, nii nagu võib muutuda terve žanri üldpilt: näitena nimetatagu olulisi nihkeid vene ja inglise ajalehekeele sõnaliigisagedustes (Сиротинина, 1968; Турыгина, 1965).

Puhtstilistiliste probleemide lahendamise kõrval saab sõnaliikide statistilist analüüsi edukalt rakendada keelte tüpoloogilisel uurimisel (Андреев, 1967, 76 jj.), tekstide automaattöötluse huvides (Белоногов, 1964; Раскина, Чепиго, 1970) ja mitmesuguste teoreetiliste lingvostatistiliste probleemide lahendamisel (Фрумкина, 1962; Кулявиня, 1969; Якубайтис, 1969; Бектаев, Лукьяненков, 1971).

Sõnaliikide statistilist analüüsi on varem teostatud ka eesti keele kohta. Võib nimetada J. Valge (1970, 1972) uurimusi sõnaliikide sagedusest eri funktsionaalsetes stiilides, M. Tiitsi, L. Veileri, K. Enniko ja S. Meimani töid sõnaliikide sagedusest A. H. Tammsaare romaani "Tõde ja õigus" I köite autorikõnes (Tiits, Veiler, 1974; Enniko, Meiman, 1975), A. Villupi (1972 ja 1974), E. Kesküla (1972) ning S. Kõiva ja E. Raadiku (1974) uurimusi adverbi esinemissagedusest eesti ilukirjandusproosas. Sõnaliikide kvantitatiivset esinemust rahvalaulus on vaadelnud E. Laugaste (1969).

Käesoleva töö eesmärgiks on lähemalt analüüsida sõnaliikide sagedusi kindlalt piiritletud allkeeles - tänapäeva ilukirjandusproosa autorikõnes, kusjuures uuritakse allkeele üldisi omadusi sõnaliikide kasutamise seisukohast ning individuaalseid erinevusi autorite vahel. Artiklis käsitletakse küsimust mitmest eri aspektist: sõnaliikide sagedusi vaadeldakse nii tekstis kui ka vastava teksti juurde kuuluvas sõnastikus, määratakse kindlaks sõnaliikide informatiivne koormus, võrreldakse tekste sõnaliigisageduste omavaheliste suhete (nn. sõnaliigi-indeksite) alusel ja arvutatakse sõnaliikidevahelised korrelatsioonid. Selline kompleksne analüüs teenib esmajoones stiiliuurimise praktilisi eesmärke, kusjuures tähelepanu on pööratud ka sõnaliikide statistilise uurimise metoodika küsimustele (materjali valik ja representatiivsus, interpretatsioon jne.).

Vaatlusmaterjalina kasutatakse 10 valimit (väljavõtet) à 5000 sõnet[+] kümne autori teosest, mis on ilmunud pärast 1960. a. Valimi üldmaht on seega 50 000 sõnet. Vaadeldavad teosed ja nende lühendid on järgmised:

[+] Sõnavarastatistika terminite ja mõistete kohta vt. Tuldava, 1971.

1. A. Beekman, Kartulikuljused. Tallinn, 1968. - A.B.
2. V. Gross, Pinginaabrid. Tallinn, 1965. - V.G.
3. A. Hint, Tuuline rand IV. Tallinn, 1966. - A.H.
4. H. Kiik, Tondiöömaja. Tallinn, 1970. - H.K.
5. J. Kross, Kolme katku vahel I. Tallinn, 1970. - J.K.
6. L. Promet, Primavera. Tallinn, 1971. - L.P.
7. V. Saar, Ukuaru. Tallinn, 1969. - V.S.
8. H. Sergo, Põgenike laev. Tallinn, 1966. - H.S.
9. M. Traat, Tants aurukatla ümber. Tallinn, 1971. - M.T.
10. E. Vetemaa, Väike romaaniraamat. Tallinn, 1968. - E.V.

Sõnaliikide sagedused teksti ja sõnastiku tasandil on arvutatud tekstidele koostatud sagedussõnastike põhjal, ülejäänud arvutustööd teostati TRÜ arvutuskeskuses.

2. MATERJALI REPRESENTATIIVSUS

Keelenähtuste statistilisel uurimisel kerkib alati probleem materjali representatiivsusest ja tulemuste usaldatavusest. Antud juhul seisneb küsimus esiteks selles, kas valimid iga autori teosest on küllalt suured, et pidada neid statistiliselt usaldatavaiks, ja teiseks, kas vaadeldavad valimid kokku representeerivad küllaldase usaldatavusega vastavat üldkogumit, s. t. tänapäeva eesti ilukirjandusproosa autorikõnet.

Valimite representatiivsuse üle otsustamine on teatavasti seotud nii kvalitatiivsete kui ka kvantitatiivsete hinnangutega (vrd. Tiit, 1971, 74 jj.). Materjali sisuline valik kuulub esmajoones selle konkreetse teadusharu kompetentsi, mille valdkonnas uurimine toimub. Sellest seisukohast lähtudes peame eesti kaasaegse ilukirjandusproosa tunnustatud autorite teoseid täiesti vastavaks uurimuse eesmärkidele. Väljavalitud kümme teost on võetud kompetentsete kirjandusteadlaste abiga koostatud nimestikust juhusliku valiku alusel. Seega on valiku printsiip õige ka statistiliselt, nimelt igale üldkogumi indiviidile on tagatud võrdne tõenäosus sattuda valimisse (Tiit, 1971, 76). Üldkogumi määramisel oleme lähtunud keelelise materjali homogeensuse nõudest ja seepärast on vaatluse all antud etapil ainult ilu-

kirjandusproosa a u t o r i k õ n e (välja on jäetud tegelaskõne, mille leksikaalsed, grammatilised jm. omadused erinevad tunduvalt autorikõnest; tegelaskõnet on mõtet uurida eraldi).

Järgnevalt peame kontrollima statistilise vaatluse teel saadud andmete usaldatavust. Selleks rakendame keelestatistikas kasutatavaid matemaatilise statistika meetodeid. Kõigepealt vaatleme üksikvalimite ja seejärel valimite seeria (üldvalimi) statistilist usaldatavust.

Valimid igast vaadeldavast teosest koosnevad meie katse puhul viiest osavalimist à 1000 sõnet. Osavalimid on võetud juhuväljavõtu printsiibil teose eri osadest terviklike lõikudena (välja jättes tegelaskõne). Selliseid 1000-sõnelisi tekstilõike nimetatakse sõnavarastatistikas "standardvalimiteks" ja need võetakse tavaliselt aluseks igasuguste sõnavarastatistiliste uurimuste läbiviimisel (vrd. Андреев, 1967, 86; Лукьяненков, 1974, 326; Якубайтис, Стурите, 1974, 45). Statistilise vaatluse teel saadud andmeid - sõnaliikide suhtelisi sagedusi - võib iga teose puhul vaadelda kui keskmisi sagedusi 1000-sõneliste tekstilõikude kohta. Statistilise usaldatavuse kontrollimisel tuleb seepärast kindlaks teha sageduste hajuvus viie tekstilõigu (osavalimi) ulatuses. Et statistilised hajuvushinnangud osutusid kõigil autoritel väga lähedasteks, toome näitena andmed ainult ühe autori - M. Traadi kohta (vt. tabel 1). Tabelis esitatakse sõnaliikide absoluutsagedused eri osavalimites (m_i), sageduste summa ($\sum m_i$), keskmine suhteline sagedus ($\bar{p}$) protsentides ning hajuvushinnangud 95%-lisel usaldusnivool: keskväärtuse absoluutne viga ehk nn. piirviga ($\varepsilon_{\bar{p}}$), keskväärtuse usalduspiirid ($\bar{p} \pm \varepsilon_{\bar{p}}$) ja suhteline viga ($\delta_{\bar{p}}$) protsentides. Peale selle hinnati iga valimirea homogeensust hii-ruudu (χ^2) abil. (Arvutuste kohta lähemalt vt. Tuldava, 1969 ja 1970.)

Nagu tabelist nähtub, on vaadeldavas teoses nimisõna keskmine suhteline sagedus 30,7 % piirveaga $\pm$ 1,7. Usalduspiirid on seega 29,0 ... 32,4 %, kusjuures keskväärtuse suhteline viga moodustab 5,5 %. Hii-ruudu väärtus (2,38) näitab, et sageduste kõikumine eri osavalimeis on juhuslikku laadi ja et kogu valimit võib nimisõnade esinemuse suhtes pidada statistiliselt homogeenseks (hii-ruudu kriitiline väärtus 5%-lisel olulisusnivool on antud katse puhul 9,49).

T a b e l 1

Sõnaliikide sagedused tekstis ja nende hajuvushinnangud M. Traadi teose "Tants aurukatla ümber" põhjal (5 osavalimit à 1000 sõnet)

Sõnaliik	Absoluutsagedused osavalimites					Kokku (Σm_i)	Keskmine $\bar{p}$(%)	Piirviga $\varepsilon_{\bar{p}}$(%)	Usalduspiirid (%)	Suhteviga $\delta_{\bar{p}}$(%)	χ^2
	m_1	m_2	m_3	m_4	m_5						
Nimisõna	297	305	324	319	293	1538	30,7	±1,7	29,0 ... 32,4	5,5	2,38
Omadussõna	72	57	64	60	55	308	6,2	±0,8	5,4 ... 7,0	12,9	2,94
Arvsõna	9	3	13	12	12	49	1,0	±0,5	0,5 ... 1,5	50,0	6,84
Asesõna	120	125	121	134	137	637	12,7	±1,0	11,7 ... 13,7	7,9	1,86
Tegusõna	221	228	223	228	233	1133	22,6	±0,6	22,0 ... 23,2	2,7	0,39
Määrsõna	166	169	149	154	155	793	15,9	±1,6	6,8 ... 10,0	6,9	7,83
Kaassõna	29	28	32	16	23	128	2,6	±0,8	1,8 ... 3,4	30,8	6,13
Sidesõna	86	79	73	76	89	403	8,1	±0,8	7,3 ... 8,9	9,9	2,25
Hüüdsõna	-	6	1	1	3	11	0,2	±0,1	0,1 ... 0,3	50,0	-
Kokku	1000	1000	1000	1000	1000	5000	100,0	-			

Seega on nimisõna esinemissagedus vaadeldavas teoses küllaltki stabiilne ja selle keskväärtus usaldusväärne, mida kinnitab ka suhtelise vea madal väärtus (5,5 %).+ Vaadeldes teiste sõnaliikide sageduste hajuvushinnanguid tabeli 1 põhjal, võime samuti konstateerida tulemuste suhteliselt head usaldatavust ja sageduste stabiilsust antud valimi ulatuses. Ainult väikese sagedusega arvsõnad, kaassõnad ja hüüdsõnad osutavad suuremat hajuvust, kuid see mõjustab vähe sõnaliikide sagedusjaotust tervikuna. Võib teha järelduse, et ilukirjandusproosa autorikõne puhul on 5000-sõnelised valimid (jaotatuna osavalimitesse 5 x 1000) sõnaliikide statistilise vaatluse jaoks küllalt representatiivsed.

Vaadeldes sõnaliigisageduste jaotumust üldvalimis, s.o. kümne autori teose ulatuses (tabel 2), võime nentida, et suuremal osal juhtudest on sõnaliikide keskmised suhtelised sagedused ($\bar{p}$) statistiliselt usaldatavad. Suhteline viga (95%-lisel usaldusnivool) ei ületa põhiliste sõnaliikide puhul 10 %, asesõna ja kaassõna sageduste viga on vastavalt 15,8 ja 12,9 ning ainult arvsõna ja hüüdsõna puhul tõuseb viga üle 20 %. Viimati nimetatud kaks sõnaliiki moodustavad aga kokku ainult ca 1 % ja nende hajuvus ei saa oluliselt mõjustada teiste sõnaliikide sagedusi. Hii-ruudu väärtused näitavad, et mõningate sõnaliikide puhul (nimisõna, asesõna ja tegusõna) on sageduste jaotumus eri valimite kaupa ebaühtlasem, kui võiks oodata homogeense valimirea korral (hii-ruudu kriitiline väärtus 5%-lisel olulisusnivool on antud juhul 16,9). See tähendab, et nimetatud sõnaliikide sagedused on eriti arvestatavad individuaalseid stiile eristavate faktoritena.

Kokkuvõttes võib väita, et meie katses osalevad kümme valimit üldmahuga 50 000 sõnet representeerivad küllaldase statistilise usaldatavusega vaadeldavat allkeelt - tänapäeva ilukirjandusproosa autorikõnet.

+ Teatavasti peetakse keelestatistilistes töödes küllaldaseks, kui suhteline viga on 10-20 %, kusjuures sõnavarastatistilistes uurimustes on lubatud suhtelise vea suurus kuni 30 % (vt. Бектаев, Пиотровский, 1974, 189).

Tabel 2

Sõnaliikide suhtelised sagedused tekstis ja nende hajuvushinnangud ilukirjandusproosa autorikõne põhjal (10 valimit à 5000 sõnet)

Sõnaliik \ Suhteline sagedus (%)	Autorid										Keskmine $\bar{p}$(%)	Piirviga $t_{\bar{p}}$(%)	Usaldus-piirid (%)	Suhteline viga $\delta_{\bar{p}}$(%)	χ^{2+}
	A.B.	V.G.	A.H.	H.K.	J.K.	L.P.	V.S.	H.S.	M.T.	A.V.					
Nimisõna	35,9	30,4	32,5	34,8	32,8	30,4	26,8	36,8	30,7	26,8	31,7	±2,6	29,1 ... 34,3	8,2	36,6**
Omadussõna	5,4	7,4	5,6	5,2	7,4	5,5	4,9	5,8	6,2	6,3	6,0	±0,6	5,4 ... 6,6	10,0	11,2
Arvsõna	0,6	1,4	1,3	1,3	1,7	0,9	0,9	1,4	1,0	0,5	1,1	±0,3	0,8 ... 1,4	27,3	12,0
Asesõna	8,3	11,7	12,9	9,3	9,1	13,8	13,8	7,9	12,7	14,4	11,4	±1,8	9,6 ... 13,2	15,8	49,3**
Tegusõna	25,0	22,2	19,1	24,2	18,3	24,0	23,5	22,0	22,6	24,0	22,5	±1,6	20,9 ... 24,1	7,1	19,7**
Määrsõna	13,9	17,4	16,6	15,0	16,5	13,9	17,3	14,8	15,9	16,3	15,8	±0,9	14,9 ... 16,7	5,7	9,5
Kaassõna	3,7	2,5	3,5	2,4	3,9	2,9	3,7	3,6	2,6	2,6	3,1	±0,4	2,7 ... 3,5	12,9	10,0
Sidesõna	7,2	6,8	8,3	7,7	10,1	8,5	9,5	7,6	8,1	9,0	8,2	±0,7	7,5 ... 8,9	8,4	11,5
Hüüdsõna	-	0,2	0,2	0,1	0,2	0,1	0,4	0,1	0,2	0,1	0,2	±0,1	0,1 ... 0,3	50,0	-
Kokku (%)	100	100	100	100	100	100	100	100	100	100	100	—			

+ Märkus: χ^2 väärtused on arvutatud sõnaliikide absoluutsageduste alusel.

3. SÕNALIIKIDE SAGEDUSED TEKSTIS

3.1. Üldised märkused

Tabelis 2 esitatakse sõnaliikide sagedused t e k s t i s . Statistilise vaatluse tulemused näitavad, et tänapäeva eesti ilukirjandusproosa autorikõnes esineb kõige rohkem nimisõnu (keskmiselt 31,7 %), seejärel tegusõnu (22,5 %), määrsõnu (5,8 %) ning asesõnu (11,4 %). Selline on sagedusjärjestus ka kõigi vaadeldavate autorite teostes, välja arvatud J. Kross, kellel on sidesõnu rohkem kui asesõnu. Erinevus nimetatud sõnaliikide keskmiste sageduste vahel on statistiliselt kindlustatud (võrreldagu keskmiste suhteliste sageduste usalduspiire, näit. määrsõnal 14,9 ... ... 16,7 % ja asesõnal 9,6 ... 13,2, s. t. usalduspiirid ei kattu, vt. tabel 2). Võib aga täheldada seda, et üksikutel autoritel on erinevused mõnede sõnaliikide sageduste vahel väiksemad kui teistel autoritel. Nii näiteks on V. Saarel ja A. Valtonil tegusõnade osakaal tekstis lähedane nimisõnade osakaalule ja vastavaid sagedusi lahutab teineteisest vaid mõni protsent. Määrsõnu esineb üldvalimis oluliselt rohkem kui asesõnu, kuid erandina kasutab L. Promet mõlemaid sõnaliike ligikaudu ühepalju (määrsõnu 13,9 ja asesõnu 13,8 %).

Nimetatud neljale kõige kasutatavamale sõnaliigile järgnevad sagedusjärjestuses sidesõna (8,3 %), omadussõna (6,0 %) ja kaassõna (3,1 %, sellest eessõnu 0,4 ja tagasõnu 2,7 %). Järjestus on samasugune kõigil autoritel, välja arvatud V. Gross, kes kasutab valimi andmeil omadussõnu rohkem kui sidesõnu (vastavalt 7,4 ja 6,8 %). Sõnaliikide sagedusjärjestuse lõpus paiknevad arvsõna (1,1 %) ja hüüdsõna (0,1 %).

Ilukirjandusproosa autorikõne sõnaliigisagedusi on mõtet võrrelda teiste eesti keele kohta teostatud statistiliste uurimustega. Tabelis 3 esitame lisaks meie katse andmetele sõnaliikide tekstisagedused A.H. Tammsaare romaani "Tõde ja õigus" I köite autorikõnes (valimi maht 60 000 sõnet; uurimuse teostasid A. Villup, K. Enniko ja S. Meiman), ajalehekeele sporditeadetes ja TASS-i sõnumites (valimi üldmaht 10 000 sõnet, vt. Valge, 1972) ja rahvalaulus (Laugaste, 1959).

Tabel 3

Sõnaliikide sagedused (%) tekstis eesti keele erinevate allkeelte põhjal

Sõnaliik	Ilukirjandusproosa, autorikõne	"Tõde ja õigus" I, autorikõne	Ajaleht: Sporditeated	Ajaleht: TASS-i sõnumid	Rahvalaul
Nimisõna	31,7	30,1	34,2	49,1	53,2
Tegusõna	22,5	23,6	22,8	19,6	24,9
Määrsõna	15,8	16,5	14,5	5,9	6,3
Asesõna	11,4	11,8	9,1	6,3	2,5
Sidesõna	8,3	10,3	6,2	6,2	0,6
Omadussõna	6,0	4,1	7,7	7,7	9,5
Kaassõna	3,1	2,8	2,4	3,4	1,8
Arvsõna	1,1	0,8	3,1	1,8	0,9
Hüüdsõna	0,1	0,0	0,0	0,0	0,3
Kokku	100,0	100,0	100,0	100,0	100,0

Võrreldes kaasaegse ilukirjandusproosa sõnaliigisagedusi A.H. Tammsaare omadega, võime nentida, et statistilised näitajad on väga lähedased. See tähendab, et tänapäeva keskmine vastab üldiselt A.H. Tammsaare sõnaliikide kasutusele. Ainult omadussõna ja sidesõna puhul võib täheldada statistiliselt olulist vahet: A.H. Tammsaare kasutab nimelt omadussõnu vähem ja sidesõnu rohkem kui kaasaegsed autorid keskmiselt. Omadussõnade protsent (4,1 %) on A. H. Tammsaarel madalam kui ühelgi teisel vaadeldaval autoril. Sidesõnu on palju ka J. Krossil (10,1 %), V. Saarel (9,5 %) ja A. Valtonil (9,0 %), kuid need autorid erinevad A. H. Tammsaarest selle poolest, et näit. J. Krossil on omadussõnade protsent kõrge (7,4 %) ja V. Saarel ning A. Valtonil on nimisõnade osakaal tekstis kõige väiksem vaadeldavate autorite seas (vastavalt 26,0 ja 26,8 %), kuna aga A.H. Tammsaarel on nimisõnu 30,1 %.

Vaadeldes meie katse taustal ajaleheteksti, näeme, et sporditeated on ilukirjandusproosa autorikõnele küllaltki lähedased, kuid siseriigi- ja välissõnumid erinevad ilukirjandusproosast oluliselt ja peamiselt selle poolest, et sõnumites on nimisõnade osakaal väga suur (49,1 %), tegusõnade ning asesõnade osakaal aga suhteliselt väike (vastavalt 19,6 ja 6,3 %).

Sõnaliigisagedused rahvalaulus peegeldavad žanri omapära, nimelt võib täheldada nimisõnade rohkust (53,2 %), kusjuures tegusõnade osatähtsus on samuti suur (24,9 %). Rahvalaulu eripäraks võib pidada ka omadussõnade suurt sa-

gedust (9,5 %). Paistab silma asesõnade, sidesõnade ja kaassõnade vähene kasutamine, võrreldes ilukirjandusproosa autorikõnega.

Huvitav on kõrvutada eesti keele andmeid vastavate uurimuste tulemustega teistest keeltest. Tabelis 4 esitame sõnaliikide sagedused viie keele ilukirjanduslikus tekstis: soome (P. Saukkoneni andmeil), ungari (Zsilka, 1973), läti (Якубайтис, 1974), vene (Клочкова, 1968) ja ukraina keel (Тищенко, 1970).

Tabel 4

Sõnaliikide esinemissagedus (protsentides) eri keelte ilukirjandusproosa tekstis+

Sõnaliik	Eesti keel	Soome keel	Ungari keel	Läti keel	Vene keel	Ukraina keel
Nimisõna	31,7	24,8	28,8	30,1	28,7	29,2
Omadussõna	6,0	9,1	6,8	5,5	7,9	6,8
Arvsõna	1,1	0,9	2,5	1,3	1,2	1,1
Asesõna	11,4	12,1	5,8	12,2	10,2	9,0
Tegusõna	22,5	29,0	21,6	23,0	18,3	19,7
Määrsõna	15,8	11,4	8,8	10,0	6,0	6,2
Kaassõna	3,1	2,3	24,6	5,9	12,2	12,5
Sidesõna	8,2	10,2		8,0	8,5	9,1
Hüüdsõna	0,2	0,2	0,1	0,3	0,0	0,0
Mitmesugused partiklid	-	-	-	3,7	7,0	6,3
Kokku	100,0	100,0	100,0	100,0	100,0	100,0

+ Eesti ja vene keele andmed pärinevad ilukirjandusproosa autorikõnest, soome keele näitajad on arvutatud kõige suurema sagedusega sõnade põhjal (mis katavad ilukirjandusproosa tekstist umbes 80 %), teiste keelte sagedused baseeruvad ilukirjandusproosa tekstidel (autorikõne ja tegelaskõne koos).

3.2. Üksikute sõnaliikide sagedused tekstis

Meie vaatluse andmeil (tabel 2) on n i m i s õ n a keskmine esinemissagedus ilukirjandusproosa autorikõnes 31,7 %. Matemaatilise statistika abil arvutatud usaldus-piirid näitavad, et hulgaliste katsete korral langeks keskmine sagedus 95-l juhul sajast piiridesse 29,1...34,3%.

Neid piire võime kokkuleppeliselt vaadelda antud allkeele "normina", pidades kõiki individuaalseid sagedusi, mis langevad neisse piiridesse, stilistiliselt neutraalseteks (mittemarkeerituteks) ja käsitades usalduspiiridest väljuvaid sagedusi oluliste individuaalsete hälvetena allkeele keskmisest.

Nimisõnade kasutamise seisukohast ületavad "normi" H. Sergo (36,8 %), A. Beekman (35,9 %) ja H. Kiik (34,8 %), "alla normi" on nimisõna sagedus A. Valtonil (26,8) ja V. Saarel (26,0).

Nimisõnade rohkust tekstis peetakse lingvostilistikas sageli "nominaalse" stiili tundemärgiks. Kuid niisuguse otsustuse tegemisel peame olema ettevaatlikud, kuna kvantitatiivne näitaja üksikult võttes ei võimalda alati adekvaatselt hinnata stiili omadusi. On vaja teostada ka kvalitatiivne analüüs ja vaadelda nimisõnade semantilisi jm. erijooni tekstides. Kvantitatiivse analüüsi täiustamiseks on otstarbekohane kindlaks teha tunnuste omavahelised suhted, näiteks võib antud juhul uurida, milliste sõnaliigisageduste arvel on paisutatud nimisõnade sagedus ja milline on tekstisageduste ja vastava sõnastiku sageduste suhe, s. t. tuleb kindlaks teha, kas tekstis esinevad nimisõnad on suure korduvusega või on tegemist nimisõnade mitmekesisusega. Neid küsimusi vaatleme artikli järgnevates osades.

Nimisõnade sagedusi võib käsitleda ka diferentseeritult, näiteks jaotades nimisõnad kahte peamisse semantilisse rühma - üldnimed ja pärisnimed. Ilukirjandusproosa autorikõnes jaotuvad nende sagedused järgmiselt:

üldnimed	26,9 % (85 % nimisõnade üldarvust)
pärisnimed	4,8 % (15 % " ")
kokku nimisõnu	31,7 % (100 %)

Üllatab pärisnimede küllaltki suur protsent tekstis. Üksikute autorite järgi kõigub pärisnimede sagedus 3,3 % (V. Gross) ja 6,0 % (H. Sergo) vahel. Võrdluseks toome andmed vene keele kohta: M. Šolohhovi romaanis "Ülesküntud uudismaa" registreeriti tekstis 3,8 % pärisnimesid, mis moodustab 16,1 % nimisõnade üldarvust (Лятина, 1968). Läti sagedussõnastiku andmeil on ilukirjanduslikes tekstides pärisnimesid 2,9 %, kuid valimisse on kaasa arvatud ka draa-

ma- ja luuletekstid, kus pärisnimesid esineb suhteliselt vähe. Registreeritud on ka pärisnimede sagedused läti keele teistes allkeeltes, nimelt ajalehekeeles - 5,8 % ja teaduslik-tehnilistes tekstides - 0,4 % (Latviešu valodas biežuma vārdnīca, 1972).

Tegusõna katab keskmiselt 22,5 % tekstist. Keskmine sagedus on arvutatud piirveaga ± 1,6, seega on usalduspiirid 20,9 ... 24,1 %. Nagu nähtub tabelist 2, on tinglikust ülempiirist suuremad väärtused A. Beekmanil (25 %) ja H. Kiigel (24,2 %). Ülempiirile lähenevad ka L. Prometi ja A. Valtoni näitajad (mõlemal 24,0 %). Huvitav on märkida, et A. Beekmanil on nii tegusõnade kui ka nimisõnade protsent oluliselt suurem allkeele keskmistest sagedustest. Kui tegusõnade sagedust seostada stiili "verbaalsusega", siis peab antud juhul nentima, et kirjaniku stiil on üheaegselt nii nominaalne kui verbaalne. Nähtavasti on aga otstarbekam nimetatud mõisteid defineerida nende omavahelise suhtena või seoses teiste sõnaliikide sagedusega (vt. allpool p. 5).

Alla "normi" kasutavad tegusõnu A. Hint (19,1 %) ja J. Kross (18,3 %). Ülejäänud autorid jäävad tegusõnade esinemissageduse poolest allkeele keskmise piiridesse.

Ka tegusõnade esinemust võib vaadelda grammatilise jm. liigituse seisukohalt. Jaotades tegusõnad käändelisteks ja pöördelisteks vormideks, saame järgmised tekstisagedused:

tegusõna käändelised vormid	8,2 %	(36 % tegusõnade üldarvust)
-"- pöördelised vormid	14,3 %	(64 % tegusõnade üldarvust)
kokku tegusõnu	22,5 %	(100 %)

Autorite seas paistab silma V. Saar tegusõna käändeliste vormide rohkuse poolest (10,4 %). Oluliselt alla keskmise esineb käändelisi vorme H. Kiigel (6,5 %) ja J. Krossil (6,4 %). Käändeliste vormide rohkus võib olla tingitud erinevatest asjaoludest (liitaegade kasutamine, seos modaalverbidega, kesksõnade esinemine täiendina jm.). Edaspidine analüüs peab välja selgitama nii ühised kui ka erinevad jooned autorite keelekasutuses. Toome siinkohal näiteks mõned tüüpilised laused V. Saare teosest "Ukuaru" (pöö-

ratagu tähelepanu tegusõnade käändeliste vormide kasutamisele):

Polnud mulle antud mahti aastaid lugeda ega arvutada, pole seda vajagi olnud, sest mu aastad on seisnud elavaina mu ümber, sirgunud, kasvanud ja nõudnud ikka rohkem ja rohkem. (lk. 8)

Niisuguses meeleolus võib üksnes mööda metsa hulkuda, sihitult hulkuda, et rahuneda ja koguda jõudu, või siis lasta kohiseda mändidel ja voolata mälestustejõel ... (lk. 11)

Minnal pole midagi maailma ja inimeste vastu, tal pole midagi varjata, peita ega häbeneda, ent ometi on tal raske tühja mannerguga läbi metsa minna: ta ei lähe viima, ta läheb tahtma ja tooma; ta ei lähe andma, vaid nuruma ja saama. (lk. 70)

Kes teab, mitu põlvkonda siin nabasid paigast venitades on kangutanud ja ubinud kivimürakaid, püüdnud neid aedadeks-aunadeks kokku veeretada. (lk. 71)

... vaatamata, mis ma panen või panemata jätan, läks ta eidega jumalaga jätma. (lk. 7)

Tegusõna pöördelisi vorme esineb eriti palju H. Kiigel (17,7 % tekstist ja 73,1 % tegusõnade üldarvust). See on tingitud peamiselt lihtaegade domineerimisest H. Kiige verbikasutuses. Keskmisest oluliselt rohkem kasutavad verbi pöördelisi vorme ka A. Beekman (16,3 %), L. Promet ja A. Valton (mõlemal 16,0 %). Kui vaadelda pöördeliste vormide protsenti tegusõnade üldarvust, siis võib esile tõsta M. Traati, kes pöördeliste vormide sageduselt (68,6 %) järgneb H. Kiigele. Teksti katmuse seisukohast jääb aga M. Traadi nimetatud näitaja keskmise piiridesse (15,5 %).

Süntaktilise funktsiooni järgi liigitatakse tegusõnad tavaliselt iseseisvateks, abi- ja modaalverbideks (Valgma, Remmel, 1968, 114). Pakub huvi tegusõna olema, mis võib esineda abiverbina, köitmena ja iseseisvas tähenduses. Meie vaatluse andmeil on tegusõna olema keskmine sagedus tekstis 5,4 %, iseseisvas tähenduses aga 2,1 % (mis moodustab ligikaudu 40 % tegusõna olema üldsagedusest). Köitme ja abiverbi funktsioonis esineb olema ligikaudu võrdse sagedusega -

vastavalt 1,7 ja 1,6 % (seega mõlemal juhul umbes 30 % üldsagedusest). Tegusõna olema osatähtsust tekstis tuleb hinnata ka selle järgi, et tekstis esinevate verbide üldsagedusest moodustab olema keskmiselt 25 %, s. t. iga neljas verb tekstis on olema (kõige sagedamini vormides on, oli, polnud).

Käändeliste ja pöördeliste vormide jaotumuse seisukohast on tegusõna olema vastavad näitajad tekstis 0,7 ja 4,7 %, s. o. 13 ja 87 % verbi olema sagedusest. Selles suhtes erineb tegusõna olema jaotumus tunduvalt verbide üldjaotumusest, nimelt on käändeliste ja pöördeliste vormide sagedused verbi kogu esinemuse taustal vastavalt 36 ja 64 % (vt. eespool). Tegusõna olema eripäraks on seega pöördeliste vormide suur ülekaal tekstis.

O m a d u s s õ n a sageduste kõikuvus ilukirjandusproosa tekstis ei ole kuigi suur: keskmiselt 5 ja 7,5 % vahel. Meie katse andmeil on kõige suuremad sagedused V. Grossil ja J. Krossil - mõlemal 7,4 %. Suhteliselt vähe kasutavad omadussõnu V. Saar (4,9 %) ja H. Kiik (5,2 %). Valimi põhjal arvutatud keskmine suhteline sagedus on 6,0 % piirveaga ± 0,6. Omadussõnade sageduse järgi tekstis on sageli hinnatud stiili "kvalitatiivsust", s. t. autori kalduvust kasutada jutustuses ja kirjelduses kvalitatiivset hinnangut. Peab aga silmas pidama, et just omadussõnade puhul osutub eriti vajalikuks kvantitatiivset uurimist seostada sisulise analüüsiga, sest omadussõnade kasutamine (mis on suhteliselt vähe seotud lause süntaktilise struktuuriga) peegeldab eriti ilmekalt autorite individuaalseid erinevusi (vrd. Шайкевич, 1968; Kraus, 1972). Omadussõnade statistilis-lingvistiline analüüs väärib kahtlemata eri uurimust.

Kolme vaadeldud sõnaliiki - nimisõnu, tegusõnu ja omadussõnu - võib pidada ilukirjandusliku teose "võtmesõnadeks", kuna need sõnaliigid kannavad peamist i n f o r m a t i i v s e t koormust ilukirjanduses (Гальперин, 1974, 128; vt. ka allpool, p. 4.2.). Nende kogusagedus ilukirjandusproosa autorikõnes ületab 70 %. Stiiliuurimise seisukohast pakuvad huvi ka ülejäänud sõnaliigid, eriti m ä ä r s õ n a d , mis katavad tekstist meie katse andmeil keskmiselt 15,8 % (usalduspiirid 14,9 ... 16,7 %). Kaks kirjanikku - V. Gross ja V. Saar - paistavad silma eriti

kõrge adverbiprotsendiga (vastavalt 17,4 ja 17,3 %), kuna aga A. Beekman ja L. Promet kasutavad määrsõnu oluliselt vähem allkeele keskmisest (mõlemad 13,9 %). Määrsõnad liigitatakse eesti keeles teatavasti kolme liiki: iseseisvad, rõhu- ja abimäärsõnad. Nende sagedusjaotumus meie valimis on järgmine:

iseseisev määrsõna	8,2 %	(52 % määrsõnade üldarvust)
rõhumäärsõna	5,6 %	(35 % " ")
abimäärsõna	2,0 %	(13 % " ")
määrsõnu kokku	15,8 %	(100 %)

Iseseisvaid määrsõnu esineb kõige rohkem V. Saarel (9,4 %, kusjuures iseseisvad määrsõnad moodustavad tema poolt kasutatud määrsõnade üldarvust 54,4 %). V. Grossil on iseseisvaid määrsõnu 9,1 % (52,4 % määrsõnade üldarvust), tema rõhumäärsõnade näitaja aga (6,8 % ehk 39 % määrsõnade üldarvust) on kõrgem kui ühelgi teisel autoril. Abimäärsõnu kasutab kõige rohkem A. Beekman (3,5 % tekstist, 25 % kõigist teoses kasutatud määrsõnadest).

Huvitav on vaadelda ka a s e s õ n a d e sageduste jaotumust eri alaliikide järgi. Kokku katavad asesõnad tekstist 11,4 %, millest 6,7 % moodustavad isikulised ja näitavad asesõnad. Järgnevas tabelis esitame andmed asesõna alaliikide sageduste kohta tekstis ja asesõnade üldarvu põhjal:

isikuline asesõna	4,4 %	(39 % asesõnade üldarvust)
näitav "	2,3 %	(20 % " ")
umbmäärane "	1,9 %	(17 % " ")
küsiv-siduv "	1,4 %	(12 % " ")
omastav "	0,7 %	(6 % " ")
enesekohane "	0,6 %	(5 % " ")
vastastikune "	0,1 %	(1 % " ")
asesõnu kokku	11,4 %	(100 %)

Autoritest kasutavad palju asesõnu A. Valton (14,4 %), L. Promet ja V. Saar (mõlemal 13,8 %). Väga vähe leidub asesõnu H. Sergo tekstis (7,9 %).

A. Valtoni kõrge asesõnade protsent sõltub peamiselt näitajate ja umbmääraste asesõnade sagedusest tekstis (vastavalt 3,7 % ja 2,5 %). Isikulisi asesõnu on A. Valtonil

4,9 % (suure sagedusega esineb sõna ta). Isikuliste asesõnade kõige kõrgema näitajaga paistavad silma L. Promet (7,1 %) ja V. Saar (6,8 %). Nende tekstis prevaleerib asesõna mina oma vormidega. Mõlemad autorid kasutavad teoses minavormi (V. Saar küll ainult osaliselt). Peab aga lisama, et L. Prometil ja V. Saarel on ka näitavate asesõnade sagedus tekstis küllaltki suur: vastavalt 2,7 ja 2,8 %. Võrdluseks võib nimetada, et A. Beekmanil on näitavaid asesõnu ainult 1,1 %.

K a a s s õ n u esineb autorikõne tekstis keskmiselt 3,1 %, neist eessõnu 0,4 % (13 % kaassõnade üldarvust) ja tagasõnu 2,7 % (87 %). Oluliselt üle "normi" esineb kaassõnu J. Krossil - 3,9 %, neist tagasõnu 3,5 %, s.o. 90 % kõigist tema poolt kasutatud kaassõnadest. Kõrge kaassõnade protsent on ka V. Saarel ja A. Beekmanil (mõlemal 3,7 %), kusjuures ees- ja tagasõnade sageduste suhe on mõlemal autoril ligikaudu võrdne (1:6). Võib arvata, et kaassõnade rohke kasutamine iseloomustab autori individuaalset stiili, näiteks kalduvust konkretiseerida ruumilisi, põhjuslikke jm. suhteid. Illustreerime seda mõne näitega J. Krossi ja A. Beekmani teostest.

Ja kõik kiikasid muudkui üles ja üles ja keerasid kaelu vasaku õla poole kõveraks. (J. K., 8).

... punakuldne kaabu tegi vahva kaare üle sinise taeva. (J. K., 7).

Ise vahtisid nad õhku ja kõigutasid tühjuse kohal koibi. (J. K., 8).

Vanker seisis parajasti kahe pärna vahel. (A.B., 5).

Küüni otsas lahtise kuuri all seisis turbahunniku kõrval looreha. (A. B., 5).

Ringi vaadanud, nägi mees kuuseheki vahelt korralikku põhukuhja. (A. B., 6).

S i d e s õ n a kvantitatiivset esinemust peavad paljud uurijad oluliseks stiilitunnuseks (Kelemen, 1964; Мистрик, 1967). Meie katse põhjal on sidesõnade keskmine sagedus autorikõnes 8,2 % usalduspiiridega 7,5 ... 8,9 %. Sidesõna alaliikide jaotumus tekstis on järgmine:

rinnastav sidesõna	5,6 %	(70 % sidesõnade üldarvust)
alistav "	2,6 %	(30 % " ")
kokku sidesõnu	8,2 %	(100 %)

Eriti palju sidesõnu esineb J. Krossil (10,1 %), V. Saarel (9,5 %) ja A. Valtonil (9,0 %). Kõigil neil autoritel domineerivad rinnastavad sidesõnad. Paistab silma rinnastava sidesõna ja kasutamine J. Krossi tekstis: valimis (5000 sõnet) esineb ja 274 korda, s. t. katab tekstist 5,5 %. Sidesõnade rohke esinemine peegeldab tavaliselt teksti süntaktilise struktuuri keerukust, näit. liitlausete või koondlausete eelistamist. Toome näiteks lause J. Krossi teosest, juhtides tähelepanu sidesõna ja kasutamisele:

Nii et ta teadis hästi, kuidas seal välja näeb, kui sada viiskümmend vahaküünalt sinab ja kõik need imelised kujud, Saalomon ja kes nad seal olid, seinavaipadelt alla vaatavad, ja kui hulk raeisandaid muskuse- ja lavendli- ja muskaadilõhna pilves kõrgel toolidel koos on ja vahivad sulle otsa ja sa pead seal midagi ütlema, ja sul on tunne, et häält pole sul ollagi, sest kõri on kuiva saepuru täis ... (J. K., 59)

J. Krossile on omane ka nn. teksti kompaktsus (Мистрик, 1967, 47), mis ilmneb selles, et lauseid seostatakse määrsõnaliselt kasutatud sidesõnaga järgneva lause alguses. Näiteks:

... täpipealt nagu isa talliuksest Kalamaja rannas. Ja väheke kõrgemalt lõhnas väravavahe vastjahvatatud nisust ... (J. K., lk. 10).

Ning selles tähelepanekus oli midagi kehutavat, teadagi, ja midagi kurvastavat kuidagi ka. Aga riiuleil ja riiulite all poepõrandal oli muidugi imeväärt kraami. (J.K., 60).

A r v s õ n a d e sagedus tekstis on keskmiselt 1,1 %. Sellest moodustavad põhiarvsõnad 0,8 % ja järgarvsõnad 0,3 %. Suhteliselt palju esineb arvsõnu J. Krossil (1,7 %), V. Grossil ja H. Sergol (mõlemal 1,4 %), näit.:

Maret kaalub viiskümmend kolm kilo. (V. G., 9).

Vanavanemad elasid veel, isa ja ema - see teeb neli.

Siis kaks sälueas isaõde ja kolm last, Juhanist nooremad. Kokku üheksa. Juhan ise ka - ongi kümme. (H.S., 10).

Paistavad silma veel ligikaudseid arve väljendavad sõnad, nagu: tosinkond (J. K.), paarkümmend, kümmekond, versta viis, tuhat korda armsam (H. S.).

H ü ü d sõ n a d e osatähtsus autorikõne tekstis on suhteliselt väike, esinevad peamiselt sellised tavalised sõnad nagu jah, ei, oh ei, noh, no, näe, ah. Kõige rohkem leidub hüüdsõnu V. Saare tekstis (22 juhtu 5000-sõnelise valimi kohta). Teistel autoritel esineb hüüdsõnu niisama suure valimi kohta keskmiselt 5 - 10 korda, A. Beekmanil ainult üks kord.

3.3. Sõnaliikide rühmitused

Sõnaliike on võimalik ühendada rühmadeks, lähtudes mitmesugustest formaalsetest või sisulistest kriteeriumidest. Sõnaliigirühmade moodustamine ja nende statistiline analüüs võib pakkuda huvi individuaalsete või funktsionaalsete stiilide uurimisel, eri keelte tüpoloogilisel kõrvutamisel, rakenduslingvistiliste probleemide lahendamisel jne. (Букович, 1972; Зореф, 1972; Кривоносов, 1973). Järgnevalt vaatleme eesti keele sõnaliike sellise rühmitamise alusel, mis võimaldab võrrelda stiile ja osaliselt kõrvutada eesti keelt mõningate teiste keeltega.

M o r f o l o o g i l i s e s t seisukohast võime eesti keele sõnad jaotada kahte suurde rühma: muutuvad sõnad ja muutumatud sõnad. Muutuvate sõnade rühma kuuluvad nimisõnad, omadussõnad, asesõnad, arvsõnad ja tegusõnad. Neid võib omakorda jaotada käändsõnadeks (neli esimesena nimetatud sõnaliiki), tegusõna käändelisteks vormideks ja pöördelisteks vormideks (pöördsõnadeks). Muutumatud sõnad oleksid seega kõigi ülejäänud sõnaliikide sõnad, kusjuures siinkohal ei võeta arvesse teatavate määrsõnade ja kaassõnade osalist muutumist kohakäänetes (näit. peal:peale:pealt). (Selliselt vaadeldakse muutuvaid ja muutumatuid sõnu ka eesti keele traditsioonilises grammatikas, vrd. Valgma, Remmel, 1968, 52).

Tabelis 3 esitati andmed üksikute sõnaliikide sagedu-ste kohta erinevate allkeelte tekstis. Koondades sõnalii-

gid morfoloogilisteks rühmadeks, saame järgmised tulemused (vt. tabel 5):

Tabel 5

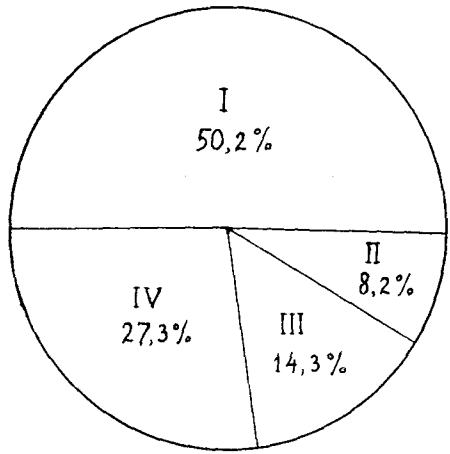
Morfoloogiliste sõnaliigirühmade sagedused (%) eesti keele eri allkeelte tekstis

Sõnaliigirühm	Tänapäeva ilukirjanduisproosa autorikõne	"Tõde ja õigus" I autorikõne	A j a l e h t		Rahvalaul
			Sport	Sõnumid	
Käändsõnad	50,2	46,8	54,1	64,9	66,1
Tegusõna käändelised vormid	8,2	22,5 23,6	22,8	19,6	24,9
Tegusõna pöördelised vormid	14,3				
Muutuvad sõnad kokku	72,7	70,4	76,9	84,5	91,0
Muutumatud sõnad kokku	27,3	29,6	23,1	15,5	9,0
Kokku	100,0	100,0	100,0	100,0	100,0

Tänapäeva ilukirjandusproosa autorikõnes domineerivad käändsõnad, moodustades tekstis 50,2 % (piirveaga ± 1,4, usalduspiirid on 48,8 ... 51,6 %). Meie vaatluse andmetel on käändsõnade protsent kõige suurem A. Hindil (52,3 %) ja H. Sergol (51,9 %), kõige väiksem aga V. Saarel (45,6 %) ja A. Valtonil (48,0 %). Muutuvaid sõnu on kaasaegsete kirjanike tekstides kokku keskmiselt 72,7 % (± 1,6). Kõige rohkem esineb neid A. Beekmanil (75,2 %), H. Kiigel (74,8 %) ja L. Prometil (74,6 %). Võrreldes tänapäeva autoreid A.H. Tammsaarega, võime nentida, et tänapäeva autorid kasutavad keskmiselt rohkem käändsõnu ja ka muutuvaid sõnu kokku, kuid mitmel üksikautoril on A.H. Tammsaarega lähedased sõnaliigirühmade sagedused. Võib teha järelduse, et ilukirjandusproosa autorikõne kvantitatiivsed näitajad on antud juhul küllaltki stabiilsed ja püsivad teatud kindlates variatsioonipiirides.

Morfoloogiliste sõnaliigirühmade jaotumust tänapäeva

ilukirjanduspoosa autorikõne tekstis aitab näitlikustada joonis 1.



Joon. 1. Morfoloogiliste sõnaliigirühmade sagedusjaotumus ilukirjanduspoosa autorikõnes (tekstis): I - käändsõnad, II - tegusõna käändelised vormid, III - tegusõna pöördelised vormid, IV - muutumatud sõnad

Pöördudes uuesti tabeli 5 poole, võrdleme ilukirjanduspoosa statistilisi näitajaid teiste allkeelte omadega. Sõnaliigirühmade sagedused toovad eriti selgelt esile erinevused allkeelte vahel. Isegi ajalehe sporditeadetes on käändsõnu tunduvalt rohkem kui ilukirjanduspoosas, ja eriti tõuseb käändsõnade osatähtsus ajalehe sõnumites (64,9 %). Kuigi sõnumites on tegusõnade protsent suhteliselt väike, saame kokkuvõttes muutuvate sõnade sageduseks tekstis 84,5 %. Muutuvate sõnade protsent on eriti suur rahvalaulus (91,0 %), selle tingivad aga teistsugused põhjused kui need, millest sõltub muutuvate sõnade osakaal ajalehekeeles. Rahvalaulus võib täheldada nii käändsõnade kui ka pöördsõnade ülikülllust, ning muutumatuid sõnu esineb vaid 9 %.

Kui tahame võrrelda eesti keele statistilisi andmeid teiste keelte vastavate näitajatega, siis on nõutav, et võrreldaks samu allkeeli või žanre. Olemasolevad andmed mõningate teiste keelte kohta pärinevad ilukirjanduspoosa tekste käsitlevatest uurimustest, milles autorikõnet ja tege-

11

laskõnet pole diferentseeritud. Tabelis 6 esitame morfoloogiliste sõnaliigirühmade sagedused ungari, läti, vene, ukraina ja rumeenia keele (Kelemen, 1964) kohta.

T a b e l 6

Morfoloogiliste sõnaliigirühmade teksti sagedused eri keeltes (%)

Sõnaliigirühm	Ungari keel	Läti keel	Vene keel	Ukraina keel	Rumeenia keel
Käändsõnad	43,9	49,1	48,0	46,1	40,6
Tegusõnad	21,6	23,0	18,3	19,7	23,0
(käändel. vormid	3,2)		(3,5)		
(pöördel. vormid	18,4)		(14,8)		
Muutuvad sõnad	65,5	72,1	66,3	65,8	63,6
Muutumatud sõnad	34,5	27,9	33,7	34,2	36,4
K o k k u	100,0	100,0	100,0	100,0	100,0

Siinjuures võib märkida, et sõnade liigitamine muutuvateks ja muutumatuteks sõnadeks on täiesti kindlapiiriline slaavi, romaani ja germaani keeltes (vrd. Кривоносов, 1973) ning võimalik ka ungari ja läti keeles (vt. näit. ungari sõnaliikide rühmitamist T. Zsilka uurimuses, 1973, lk. 101 jj.). Toodud andmeid võime võrrelda eesti keelega, silmas pidades teatavat erinevust allkeeltes, s. t. tegelaskõne mõju võõrkeelse ilukirjandusproosa kvantitatiivsetele karakteristikutele. Võib aga arvata, et antud juhul autorikõne ja tegelaskõne tasakaalustuvad: ühelt poolt esineb tegelaskõnes vähem käändsõnu, teiselt poolt rohkem tegusõnu, seega peaks muutuvaid sõnu olema mõlemas ligikaudu võrdselt. Muutuvate sõnade osatähtsuselt tekstis on eesti keel lähedane läti keelele, ületades teised vaadeldavad keeled tunduvalt.

Väärib tähelepanu sõnade rühmitamine s e m a n t i l i s - s ü n t a k t i l i s e printsiibi järgi, mille kohaselt moodustuvad kaks suurt rühma: "autosemantilised" ehk täistähenduslikud sõnad (vene k. полнозначные, знаме-

знательные слова; saksa k. Vollwörter) ja "sünsemantilised" sõnad, mida võib tinglikult nimetada "abisõnadeks" või "struktuurisõnadeks" (vene k. строевые, структурные, служебные слова; saksa k. Leerwörter, Strukturwörter). Niisuguse sõnade liigitamine kaheks suureks rühmaks on tuntud juba Aristotelese ajast ja selle alusel on korduvalt uuritud stiile ja võrreldud keeli tüpoloogilisest seisukohast. Täpse piiri tõmbamine täistähenduslike ja mittetäistähenduslike sõnade vahele on esile kutsunud vaidlusi. Keelestatistikas on välja kujunenud kaks skeemi, millest lähtuvad kvantitatiivsed uurimused. P. Guiraud' (1959) järgi peetakse täistähenduslikeks sõnadeks ainult nimisõnu, omadussõnu, tegusõnu ja määrsõnu (eesti keele seisukohast iseseisvaid määrsõnu). On tuntud teinegi liigitus (näit. Těšitelová, 1972), mis arvab täistähenduslike sõnade hulka ka arvsõnad ja asesõnad. Ka eesti keele grammatikas eristatakse täistähenduslikke sõnu, (Mihkla jt., 1974, 12), lähtudes nii semantilisest kui ka süntaktilisest printsiibist.

Käesolevas töös võtame aluseks mõlemad skeemid: I - täistähenduslikud sõnad kitsamas mõttes (nimisõnad, omadussõnad, tegusõnad, iseseisvad määrsõnad) ja ülejäänud sõnad; II - täistähenduslikud sõnad laiemas mõttes (kaasa arvatud arvsõnad ja asesõnad) ning abisõnad.[+] Sellise jaotuse alusel saame eri autorite kohta järgmised tekstisagedused (protsentides):

	A.B.	V.G.	A.H.	H.K.	J.K.	L.P.	V.S.	H.S.	M.T.	A.V.
I -	73,9	69,0	65,5	71,3	67,3	66,3	63,8	72,6	67,9	65,8
II -	82,8	82,1	79,7	81,9	78,1	81,0	78,5	81,9	81,6	80,7

Esimese skeemi järgi esineb autorikõne tekstis keskmiselt 68,3 % täistähenduslikke sõnu (usalduspiirid 65,9 ... ... 70,7 %). Variatsiooniulatus on suur, nimelt kõiguvad sagedused 63,8 % (V. Saar) ja 73,9 % (A. Beekman) vahel. See tähendab, et täistähenduslike sõnade protsent on esimese skeemi järgi individuaalseid stiile eristav faktor. Teise skeemi järgi, s. t. arvestades täistähenduslike sõnade hulka

[+] Rangelt võttes tuleks täistähenduslike sõnade hulgast välja arvata tegusõna olema abiverbi funktsioonis (mis moodustab keskmiselt 1,6 % tekstist). Käesoleval juhul arvestame verbi olema sagedust tervikuna, et andmed oleksid paremini võrreldavad teiste keeltega.

ka arv- ja asesõnad, on autorikõne keskmine 80,8 %, kusjuures kõikuvus on väike (usalduspiirid: 79,7 ... 81,9 %, suhteline viga ainult 1,4 %!). Juba varem on juhitud tähelepanu sellele, et täistähenduslike sõnade protsent tekstis on teise skeemi järgi ühe keele ulatuses väga stabiilne (näit. tšehhi keeles kõigub täistähenduslike sõnade sagedus 80% ümber, vt. Těšitelová, 1972). Nimetatud tunnus võimaldab keeli uurida tüpoloogia seisukohast. Esitame alljärgnevalt kõrvutavad andmed täistähenduslike sõnade sageduse kohta kahe skeemi järgi eesti keeles ja mõnedes teistes keeltes.

	I	II
Eesti keel (ilukirjandusproosa autorikõne)	68,3 %	80,8 %
Ungari keel (ilukirjandusproosa)	69,5 %	78,9 %
Läti keel (ilukirjandusproosa)	68,6 %	82,1 %
Vene keel (ilukirjandusproosa autorikõne)	60,9 %	72,3 %
Ukraina keel (ilukirjandusproosa)	61,9 %	72,0 %
Saksa keel (ilukirjandusproosa autorikõne)+	55,2 %	71,9 %

Paistab silma täistähenduslike sõnade suhteliselt kõrge sagedus eesti, läti ja ungari keeles (nii I kui ka II skeemi järgi).

4. SÕNASTIKU- JA TEKSTISAGEDUSED; FUNKTSIONAALNE JA INFORMATIIVNE KOORMUS

4.1. Sõnaliikide sagedused sõnastikus

Sõnaliikide statistilisel uurimisel pakub huvi nende esinemissagedus vaadeldavate tekstide juurde kuuluvais s õ n a s t i k e s . Käesolevas töös on sõnastikud koostatud 7 teksti kohta sõnavormide tasandil. Sõnaliikide suhtelised sagedused ja hajuvushinnangud esitame tabelis 7. Nagu näha, on ka sõnastikus nimisõnad sageduselt esikohal (keskmiselt 44,2 %), teisel kohal on tegusõnad (26,6 %) ja kolmandal kohal määrsõnad (10,7 %). Tuleb aga silmas pidada, et meie näites on sagedused arvestatud sõnavormide sõnastiku põhjal. Võrdluseks võib tuua andmed uurimusest

+ Saksa keele andmed on võetud L. Pilleri artiklist (Piller, 1971). Teiste keelte kohta kasutatakse varem nimetatud uurimuste andmeid.

Tabel 7

Sõnaliikide suhteline sagedus sõnastikus (sõnavormide tasandil) eesti ilukirjandusproosa autorikõne põhjal (7 valimit à 5000 sõnet)

Sõnaliik \ Suhteline sagedus (%)	Autorid							Keskmine $\bar{p}$ (%)	Piirviga $\varepsilon_{\bar{p}}$ (%)	Usalduspiirid (%)	Suhteline viga $\delta_{\bar{p}}$ (%)
	A.B.	V.G.	H.K.	J.K.	L.P.	V.S.	H.S.				
Nimisõna	40,3	42,7	45,9	45,9	44,9	39,3	50,2	44,2	± 3,5	40,7 ... 47,7	7,9
Omadussõna	8,4	12,0	8,3	12,2	8,5	8,7	8,3	9,5	± 1,7	7,8 ... 11,2	17,9
Arvsõna	0,6	1,6	1,6	1,5	1,0	1,1	1,7	1,3	± 0,4	0,9 ... 1,7	30,8
Asesõna	3,5	5,0	4,9	3,5	4,9	5,6	3,8	4,4	± 0,8	3,6 ... 5,2	17,8
Tegusõna	32,6	24,5	27,1	20,1	27,6	30,9	23,7	26,6	± 4,0	22,6 ... 30,6	15,0
Määrsõna	11,3	11,1	9,1	13,1	10,0	10,6	9,3	10,7	± 1,3	9,4 ... 12,0	12,1
Kaassõna	2,2	2,0	2,1	2,5	2,1	2,5	2,0	2,2	± 0,2	2,0 ... 2,4	9,1
Sidesõna	1,0	0,8	0,7	0,9	0,9	1,0	0,8	0,9	± 0,1	0,8 ... 1,0	11,1
Hüüdsõna	0,1	0,3	0,3	0,3	0,1	0,3	0,2	0,2	± 0,1	0,1 ... 0,3	50,0
Kokku (%)	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	-	-	-
Sõnastiku maht (V)	2870	2783	2643	2821	2721	2447	2892	keskm. 2740	-	-	-

A.H. Tammsaare romaani "Tõde ja õigus" I köite autorikõne kohta (Enniko, Meiman, 1975), kus sõnaliikide sõnastikusagedused on antud sõnade (lekseemide) tasandil: nimisõnu - 43,4 %, tegusõnu - 22,8 %, omadussõnu - 8,6 %, arvsõnu - 0,9 %, asesõnu - 2,1 %, määrsõnu - 17,2 %, kaassõnu - 3,5%, sidesõnu - 1,4 %, hüüdsõnu - 0,1 %. Lekseemide tasandil väheneb mõnevõrra muutuvate sõnade osakaal, kuna aga muutumatute sõnade osatähtsus suureneb (näiteks on määrsõnu sõnavormide sõnastikus keskmiselt 10,7 %, A.H. Tammsaare lekseemide sõnastikus aga 17,3 %).

Morfoloogilise rühmituse järgi jagunevad sagedused ilukirjandusproosa autorikõne sõnastikus sõnavormide ja lekseemide tasandil järgmiselt:

	Sõnavormid (7 autorit)		Lekseemid (A.H. Tammsaare)
Käändsõnad	59,4 %		54,0 %
Tegusõna käändel. vormid	12,0 %	26,5 %	
Tegusõna pöördel. vormid	14,6 %		23,5 %
Muutuvad sõnad (kokku)	86,0 %		77,5 %
Muutumatud sõnad (kokku)	14,0 %		22,5 %

Täistähenduslikke sõnu (sõnavorme) on kaasaegse ilukirjandusproosa autorikõne sõnastikus I skeemi järgi (s.o. nimisõnad, omadussõnad, tegusõnad ja iseseisvad määrsõnad) 87,6 % ning II skeemi järgi (kaasa arvatud ase- ja arvsõnad) - 93,3 %. Lekseemide tasandil on vastavad sagedused A.H. Tammsaare sõnastikus: 83,0 % ja 86,4 %.

4.2. Sõnaliikide funktsionaalne ja informatiivne koormus

Sõnaliikide funktsionaalset koormust võib määritleda mitmel viisil. Kõige lihtsam on arvestada funktsionaalse koormusena vastava sõnaliigi suhtelist sagedust, s. o. suhtelist osatähtsust tekstis või sõnastikus. Selline vaatlusviis ei too aga iseenesest esile mõningaid olulisi sisemisi seaduspärasusi, mis on omased sõnaliikide funktsioneerimisele tekstis. On otstarbekohane vaadelda sõnaliikide esinemust teksti ja sõnastiku sageduste suhtena, mis annab meile vastavasse sõnaliiki kuuluvate sõnade k e s k m i s e

k o r d u v u s e tekstis:

$$KK = t/s,$$

kus KK tähistab korduvus- ehk iteratsiooni-indeksit, t - tekstisagedust ja s - sõnastikusagedust (antud juhul sõnavormide tasandil). Tabelis 8 on toodud andmed eri sõnaliikide keskmise korduvuse kohta kaasaegse ilukirjandusproosa autorikõne tekstis. Kõige suurema keskmise korduvusega on sidesõnad: igaüks neist kordub keskmiselt 17 korda 5000-sõnelises tekstis. Selle põhjal võib väita, et sidesõnad on tekstis kõige suurema koormusega. Kõige väiksema korduvusega on omadussõnad (KK = 1,1). Kui sõnade korduvus tekstis on väike, siis tähendab see ühtlasi seda, et nende mitmekesisus on suur.

Pakub huvi vaadelda sõnaliikide korduvust ka morfoloogiliste rühmituste järgi. Käändsõnad korduvad 5000-sõnelises valimist keskmiselt 1,5 korda, tegusõna käändelised vormid 1,3 ja pöördelised vormid 1,8 korda. Muutuvad sõnad (sõnavormid) korduvad keskmiselt 1,6 korda, muutumatud sõnad aga 3,6 korda.

Sõnaliikide funktsionaalse koormuse mõõtmine korduvusindeksi alusel võimaldab kindlaks teha olulisi erinevusi autorite individuaalsetes stiilides. Vaatleme näiteks nimisõnade esinemust tekstis ja sõnastikus ning iga nimisõna keskmist korduvust seitsme autori tekstis.

	t	s	KK = t/s
A. Beekman	1796	1157	1,55
V. Gross	1519	1189	1,28
H. Kiik	1739	1214	1,43
J. Kross	1642	1295	1,27
L. Promet	1519	1222	1,24
V. Saar	1298	961	1,35
H. Sergo	1842	1452	1,27

A. Beekmani ja H. Sergo tekstis esineb palju nimisõnu, kuid korduvusindeks näitab olulist erinevust nende kasutuses. A. Beekmanil on nimisõnade korduvus suurem (keskmiselt 1,55) ja seega nende mitmekesisus väiksem, H. Sergol aga korduvad nimisõnad suhteliselt vähe (1,27), s. t. nende valik on suurem. Teiste sõnaliikide osas võib aga olukord olla vastupidine (nimetatud kirjanike puhul täheldamegi

Tabel 8

Sõnaliikide keskmine esinemissagedus tekstis ja sõnastikus, keskmine korduvus ja informatiivne koormus

Sõnaliik	Keskm. absol. sagedus		Keskmine korduvus (t/s)	Keskm. suhtel. sagedus (%)		Informatiivne koormus (p_s/p_t)
	tekstis (t)	sõnastikus (s)		tekstis (p_t)	sõnastikus (p_s)	
Nimisõna	1585	1213	1,3	31,7	44,2	1,4
Omadussõna	300	260	1,1	6,0	9,5	1,6
Arvsõna	55	36	1,5	1,1	1,3	1,1
Asesõna	570	121	4,4	11,4	4,4	0,4
Tegusõna	1125	727	1,5	22,5	26,6	1,2
Määrsõna	790	292	2,7	15,8	10,7	0,7
Kaassõna	155	61	2,5	3,1	2,2	0,7
Sidesõna	410	24	17,0	8,2	0,9	0,1
Hüüdsõna	10	6	1,7	0,2	0,2	1,0
Kokku	5000	2740	1,8	100,0	100,0	-

seda tegu-, määr- ja omadussõnade esinemuses). Esitame alljärgnevalt omadussõnade, tegusõnade ja määrsõnade korduvusindeksi väärtused vaadeldava seitsme valimi lõikes:

	A.B.	V.G.	H.K.	J.K.	L.P.	V.S.	H.S.
Omadussõnu	1,12	1,11	1,18	1,08	1,18	1,17	1,21
Tegusõnu	1,34	1,63	1,69	1,61	1,60	1,55	1,61
Määrsõna	2,14	2,82	3,12	2,24	2,57	3,34	2,76

Omadussõnade korduvus on kõige väiksem J. Kroosil (KK= = 1,08), tegu- ja määrsõnade korduvusindeksid on aga kõige väiksemad A. Beekmanil (vastavalt 1,34 ja 2,14). Suured indeksite väärtused (ja väiksem mitmekesisus) iseloomustavad H. Sergo omadussõnu (1,21), H. Kiige tegusõnu (1,69) ja V. Saare määrsõnu (3,34).

Korduvus- ehk iteratsiooni-indeksi kasutamine funktsionaalse koormuse mõõduna on ülevaatlik ja sisuliselt arusaadav. Indeksi puuduseks on asjaolu, et eri andmete võrdlemisel peab lähtuma ühesuurustest valimitest. Indeksi väärtus sõltub täielikult valimi mahust. Seepärast on mõtet otsida teisi võimalusi sõnaliikide tekstikoormuse mõõtmiseks, et saaks võrrelda eri uurimuste andmeid. Normeerides absoluutsagedused sel lihtsal viisil, et arvutame vastavad suhtelised sagedused, võime funktsionaalset koormust mõõta teksti ja sõnastiku suhteliste sageduste suhtena. Uurimused on näidanud, et sõnaliikide suhtelised sagedused jäävad valimi mahu (teksti pikkuse) suurenedes enamvähem konstantseteks nii sõnastiku kui ka teksti tasandil (sõnastiku tasandil vähenevad küll selliste sõnaliikide nagu ase-, kaas- ja sidesõnade suhtelised sagedused, kuid nende osatähtsus sõnastikus on väike ja nad ei avalda olulist mõju arvutustele). Sõnaliikide funktsionaalset koormust võib seega määrata teksti ja sõnastiku suhteliste sageduste jagatise abil:

$$FK = p_t/p_s ,$$

kus FK on normeeritud funktsionaalse koormuse mõõt, p_t - - sõnaliigi suhteline sagedus tekstis ja p_s - sõnaliigi suhteline sagedus sõnastikus.

Kirjeldatud suhet võib vaadelda ka ümberpöördult, s. o. jagades sõnastikusageduse tekstisagedusega. Tuginedes informatsiooniteooria põhimõtetele, võib sel viisil saadavat suhet sisuliselt interpreteerida kui sõnaliigi i n f o r -

matiivset koormust (IK):

$$IK = p_s/p_t .$$

Mida suurem on IK väärtus, seda suurem on sõnaliigi suhteline osakaal sõnastikus ja järelikult on iga esinemine tekstis suhteliselt haruldasem ning ühtlasi informatiivsem. Informatiivne koormus kirjeldatud mõttes on lähedane nn. mitmekesisuse indeksile TTR ("type-token-ratio", vt. Tuldava, 1971, 220) ja väljendab samuti sõnavara mitmekesisust või "rikkust", antud juhul sõnaliigi kohta terviklikult. Võrreldes TTR-indeksiga sõltub aga IK väga vähesel määral valimi suurusest ja seepärast on võimalik IK abil võrrelda erinevatest katsetest ja erinevate keelte põhjal saadud tulemusi (võrreldavate valimite mahud ei tohi siiski olla liiga erinevad).

Tabelis 8 on toodud informatiivse koormuse (IK) väärtused eri sõnaliikide kohta meie katse materjali põhjal. Kõrvutatakse keskmisi suhtelisi sagedusi sõnastikus ja tekstis. Informatiivse koormuse alusel võib sõnaliigid jagada kahte rühma: sõnaliigid, mille IK väärtus on üle 1, ja sõnaliigid, mille IK väärtus on 1 või alla selle. Esimesse rühma kuuluvad eelkõige omadussõnad (IK = 1,6), nimisõnad (IK = 1,4) ja tegusõnad (IK = 1,2). Suhteliselt kõrget informatiivset koormust tekstis omavad meie katse andmeil ka arvsõnad (IK = 1,1). Määrsõnade informatiivne koormus tekstis on kokkuvõttes suhteliselt madal. Vaadeldes määrsõna eri liike omaette, saame järgmised IK väärtused: iseseisvad määrsõnad - 0,9, rõhumäärsõnad - 0,4 ja abimäärsõnad - 0,5.

Sõnaliikide informatiivset koormust võib vaadelda diferentseeritult ka morfoloogiliste rühmade kaupa. Ilukirjandusproosa autorikõne kohta saame järgmised tulemused: käändsõnad - 1,2, tegusõna käändelised vormid - 1,5, tegusõna pöördelised vormid - 1,0; muutuvad sõnad kokku - 1,2, muutumatud sõnad - 0,5. Seega osutuvad kõige informatiivsemaks tegusõna käändelised vormid.

Informatiivsuse indeks IK peegeldab täiesti ootuspäraselt ka seda seika, et täistähenduslikel sõnadel on tekstis märgatavalt suurem informatiivne koormus kui abisõnadel. Nii saame näiteks I skeemi alusel (vt. 3.3.) täistähenduslike sõnade IK väärtuseks 1,3: ülejäänud sõnade IK väärtuseks (kokku) aga ainult 0,4. II skeemi järgi (arvestades täis-

tähenduslike sõnade hulka ka arv- ja asesõnad) on vastavad väärtused 1,1 ja 0,3.

5. SÕNALIIGI-INDEKSID

Funktsionaalsete ja individuaalsete stiilide kvantitatiivsel uurimisel on paljudes töödes lähtutud sõnaliigisageduste omavahelistest suhetest. Eriti huvipakkuvad on niisugused sõnaliikide suhted, mida saab sisuliselt interpreteerida. Nii näiteks on juba ammu tuntud ja stilostatistikas ning psühholingvistilistes uurimustes kasutatud nn. Busemanni koefitsient, mis arvutatakse omadussõna ja tegusõna tekstisageduste suhtena (vt. Busemann, 1925 ja 1948; Boder, 1940; Antosch, 1969; Fischer, 1965; Zsilka, 1974). Omadussõna ja tegusõna vastandamine põhineb tähelepanekul, et mõlemad sõnaliigid on oma semantilise funktsiooni poolest predikatiivse toimega, s. t. nad ütlevad midagi uut kõnesoleva objekti kohta (vrd. Гальперин, 1974, 164–165). Olenevalt individuaalsest stiilist kasutavad ühed autorid esemete ja nähtuste kirjeldamisel rohkem adjektiivset ("kvalitatiivset") hinnangut, teised aga eelistavad verbaalset väljendusviisi. Seepärast saab omadussõnade ja tegusõnade sageduste suhet vaadelda kui stiili "kvalitatiivsuse" või selle vastandi – "verbaalsuse" või "tegevuslikkuse" hinnangut (vrd. vene keele stilistikas kasutusel olevaid termineid качественность ja действенность vt. näit. Папина, Иванова-Маркова, 1974).

Stilostatistikas on laialdast kasutamist leidnud veel mitmesugused muud sõnaliigisageduste suhted, mida saab sisuliselt tõlgendada ja mis on uurimuste tulemusena osutunud individuaalseid või funktsionaalseid stiile eristavateks ja diagnoosivateks faktoriteks (vt. näit. Мистрик, 1967; Головин, 1971; Muller, 1967; Renský, 1965; Těšitelová, 1972).

Käesolevas töös vaatleme mõningaid tuntuimaid sõnaliigi-indekseid eesti ilukirjandusproosa autorikõne põhjal. Tulemused esitame kokkuvõtlikul kujul tabelis 9. Alljärgnevalt kommenteerime lühidalt tabeli andmeid.

1. Esimese sõnaliigi-indeksina on tabelis toodud omadussõnade ja verbi pöördeliste vormide sageduste suhe: A/V_p.

See suhe ("Busemanni koefitsient") määrab stiili k v a l i t a t i i v s u s e astme, s. t. adjektiivide abil väljendatud kvalitatiivse hinnangu osatähtsuse, võrreldes verbaalse väljendusviisiga. Funktsionaalsete stiilide võrdlemisel on kvalitatiivsuse indeks alati suurem publitsistlikus ja teaduslikus tekstis ja väiksem ilukirjanduses, eriti draamas. Allkeelte piirides võib omakorda täheldada individuaalseid erinevusi kvalitatiivsuse määras. Meie katse andmeil on kümne valimi keskmine indeksi väärtus 0,43. Arvestades usalduspiire 95%-lisel usaldusnivool (0,43 ± 0,08), võib nentida, et kahel autoril (J. Krossil ja V. Grossil) on indeksi väärtus märgatavalt üle keskmise. Antud juhul korreleeruvad kvalitatiivsuse hinnang ja omadussõnade sagedus tekstis (vt. tabel 2). Kuid alati ei tarvitse see nii olla. Näiteks A. Hindil ja L. Prometil on omadussõnade protsent tekstis ligikaudselt võrdne: vastavalt 5,6 ja 5,5 %. Tegusõna pöördelisi vorme on aga A. Hindil 11,1 % ja L. Prometil 16,0 %. Seega erinevad kvalitatiivsuse hinnangud neil autoritel tublisti: A. Hindi stiil on "kvalitatiivsem" ülalkirjeldatud mõttes (indeksi väärtus - 0,50), L. Prometi stiil on aga "verbaalsem" (indeksi väärtus - 0,34).

2. Omadussõnade keskmist sagedust iga nimisõna kohta väljendab nn. modifikatsiooni- ehk a d j e k t i i v s u s e indeks A/S. Ilukirjanduse proosa autorikõne keskmine on 0,19. Seega esineb tekstis keskmiselt üks omadussõna viie nimisõna kohta. Võrreldes teiste keelte analoogiliste andmetega on see näitaja eesti keeles suhteliselt madal. Näiteks ungari ilukirjanduse proosas on vastav indeksi väärtus 0,33 (Zsilka, 1973, 118), vene keeles E. Šteinfeldti sagedussõnastiku andmeil (Штейнфельдт, 1963, 31) - 0,31. Eesti keele ajalehetekstide sporditeadetes on omadussõnade ja nimisõnade suhe 0,22 ja välisteadetes 0,16 (Valge, 1972, 32-33). A.H. Tammsaare "Tõde ja õiguse" autorikõnes on adjektiivsuse indeksi väärtuseks 0,13 (Enniko, Meiman, 1975). Käesolevas töös vaadeldavate autorite seas on suhteliselt kõrged indeksi väärtused V. Grossil ja A. Valtonil (kummalgi 0,24) ning J. Krossil (0,23).

Tabel 9

Sõnaliigi-indeksid ilukirjandusproosa autorikõne põhjal
(10 valimit à 5000 sõnet)

Autor / Indeks	A.B.	V.G.	A.H.	H.K.	J.K.	L.P.	V.S.	H.S.	M.T.	A.V.	Keskm. $\bar{x}$	Piirviga $\varepsilon_{\bar{x}}$	Suhtel. viga $\delta_{\bar{x}}$(%)
A/V_p [+] kvalitatiivsus	0,33	0,58	0,50	0,29	0,63	0,34	0,38	0,44	0,40	0,40	0,43	±0,08	18,6
A/S adjektiivsus	0,15	0,24	0,17	0,15	0,23	0,18	0,19	0,16	0,20	0,24	0,19	±0,02	10,5
S/V substantiivsus	1,44	1,37	1,70	1,44	1,80	1,27	1,11	1,68	1,36	1,12	1,43	±0,17	11,9
$V_p/(S+P+A)$ verbaalsus	0,33	0,26	0,22	0,36	0,24	0,32	0,29	0,26	0,31	0,34	0,29	±0,03	10,3
P/S pronominaalsus	0,23	0,38	0,40	0,27	0,28	0,45	0,53	0,22	0,41	0,54	0,37	±0,08	21,6
D_i/V_p adverbiaalsus	0,47	0,71	0,75	0,40	0,74	0,41	0,72	0,61	0,54	0,55	0,59	±0,10	16,9

[+] Lühendite tähendused: A - adjektiiv, V - verb, V_p - verbi pöördeline vorm, S - substantiiv, P - pronoomen, D_i - iseseisev adverb

3. Nominaalsust ja verbaalsust on stilostatistilistes töödes sageli hinnatud nimisõnade ja tegusõnade sageduste suhtena. Sellist suhet peetakse tähtsaks kvantitatiivseks karakteristikuks funktsionaalsete stiilide kõrvutaval uurimisel (Кожина, 1972, 140). Mõnede uurijate arvates toob nimisõnade ja tegusõnade sageduste suhe esile erinevused stiilide "staatilisuse" ja "dünaamilisuse" vahel (Zsilka, 1967, 156). Suhet võib esitada mõlemapoolselt, s. o. nii nimisõnade suhtena tegusõnadesse kui ka ümberpöördult. Käesoleva materjali põhjal arvutame nimisõnade sageduse jagatise tegusõnade sagedusega ja nimetame selle s u b s t a n t i i v s u s e indeksiks. Indeksi keskmine on autorikõne valimite põhjal 1,43. Ka teistes keeltes on ilukirjandusproosa substantiivsuse hinnang lähedane eesti keelele: ungari keeles - 1,45, tšehhi keeles - 1,44 (T. Zsilka andmeil), ukraina keeles - 1,49 (Ткаченко, 1970, 217 jj.). Eesti ajalehekeele sporditeated on lähedased ilukirjandusproosa autorikõnele (indeksi väärtus - 1,51), kuid välisteadetes on substantiivsuse aste märksa kõrgem - 2,51 (Valge, 1972, 32-33). A.H. Tammsaare autorikõnes on verbide ja nimisõnade sageduste suhe 1:1,22.

Kaasaegse ilukirjandusproosa autorikõne keskmise ületavad oluliselt J. Kross, A. Hint ja H. Sergo. Väga madalat substantiivsust võib nentida V. Saare ja A. Valtoni (vaadeldavates) teostes - indeksi väärtused on vastavalt 1,11 ja 1,12.

4. Tuntud nõukogude keelestatistik B. Golovin on teinud ettepaneku stiili verbaalsust ja ühtlasi "tegevuslikkust" ning "dünaamilisust" määratleda verbisageduste suhtena noomenite (nimisõnade, omadussõnade ja asesõnade) sagedustesse tekstis (Головин, 1971, 147). Kasutades seda v e r b a a l s u s e hinnangut, saame autorikõne keskmiseks 0,29, kusjuures keskmisest oluliselt kõrgemat verbaalsust võib täheldada eriti H. Kiigel (0,36) ja A. Valtonil (0,34), samuti A. Beekmanil (0,33) ja L. Prometil (0,32).

5. Asesõnade ja nimisõnade suhet (P/S) võib vaadelda p r o n o m i n a a l s u s e hinnanguna. See näitaja on tundlik individuaalsete erinevuste suhtes ja võimaldab avastada ka sellise iseärasuse nagu minavormi kasutamine,

juhul kui asesõnadest arvestada ainult isikulisi asesõnu indeksi komponendina. Meie katse andmeil on kõrge pronominaalsuse aste L. Prometi ja V. Saare (kes kasutavad minavormi) ning A. Valtoni stiilil.

6. Iseseisvate määrsõnade ja tegusõna pöördeliste vormide sageduste suhe näitab, kui palju keskmiselt kasutatakse täistähenduslikke määrsõnu iga verbi (pöördelise vormi) kohta tekstis. A d v e r b i a a l s u s e indeksi väärtused kõiguvad vaadeldavates valimites 0,40 (H. Kiik) ja 0,75 (A. Hint) vahel. Indeksi keskmise väärtuse ületavad oluliselt ka J. Krossi, V. Saare ja V.Grossi näitajad (vastavalt 0,74; 0,72; 0,71). Adverbiaalsuse suhteliselt kõrget astet võib täheldada ka A.H. Tammsaare autorikõnes (0,70).

Kokkuvõttes tuleb nentida, et sõnaliikide sageduste suhteid väljendavad sõnaliigi-indeksid aitavad täpsustada sõnaliikide funktsioone tekstis ja võimaldavad sügavamalt läheneda sõnaliikide stilostatistilisele analüüsile.

6. SÕNALIIKIDEVAHELISED KORRELATSIOONID

Sõnaliikide kasutamist tekstis on mõtet vaadelda kui süsteemset nähtust, s. o. elementide omavaheliste seoste ja sõltuvuste alusel. Eespool kirjeldatud sagedussuhete uurimise kõrval võib sõnaliikidevahelisi seoseid kindlaks teha ja mõõta statistilise korrelatsioonanalüüsi abil. Ilukirjandusproosa autorikõne kümne valimi põhjal arvutatud korrelatsioonid (Pearsoni-Bravais' meetodi järgi, vt. lähemalt: Tuldava, 1972) esitatakse tabelis 10.

Korrelatsioonikordajate (r) kriitilised väärtused on 10%-lisel olulisusnivool 0,54 ja 5%-lisel olulisusnivool 0,63. Peab märkima, et statistiliselt olulisi korrelatsioone ei ole meie katse puhul kuigi palju. See tähendab, et sõnaliike kasutatakse tekstis suhteliselt sõltumatult üksteisest. Siiski võib juhtida tähelepanu mõningatele olulistele seostele, näiteks tugevale negatiivsele korrelatsioonile nimisõna ja asesõna sageduste vahel (-0,89), mõõdukale negatiivsele korrelatsioonile nimisõna ja määrsõna sageduste vahel (-0,60) ning negatiivsele seosele tegusõna ja oma-

dussõna vahel (-0,54). Nimisõna ja asesõna vaheline negatiivne seos on sisuliselt hästi seletatav: need sõnaliigid konkureerivad esemete ja nähtuste nimetamisel kõnes. Tegusõna ja omadussõna seos tuleneb "kvalitatiivsuse" ja "tegevuslikkuse" (verbaalsuse) konkurentsist esemete ja nähtuste kirjeldamisel (vt. eespool, p. 5). Raskem on leida otsest seletust nimisõna ja määrsõna vahelisele seosele, samuti tegusõna ja arvsõna sageduste negatiivsele korrelatsioonile (r = -0,75). Ilmselt on siin tegemist vastastikuste korrelatsioonide läbipõimumisega, mida aitaks lahti mõtestada osa- ja mitmese korrelatsiooni arvutamine (lähemalt vt. Tiit, 1972, 93 jj.).

T a b e l 10

Korrelatsioonid sõnaliikide tekstisageduste vahel

Sõnaliik	2.	3.	4.	5.	6.	7.	8.
1. Nimisõna	0,05	-0,89	-0,12	-0,60	0,24	-0,49	0,40
2. Omadussõna		-0,10	-0,54	0,43	-0,13	0,06	0,46
3. Asesõna			0,11	0,41	-0,36	0,33	-0,44
4. Tegusõna				-0,48	-0,42	-0,39	-0,75
5. Määrsõna					-0,02	0,32	0,30
6. Kaassõna						0,40	0,18
7. Sidesõna							0,05
8. Arvsõna							-

Korrelatsioonimaatriksist lähtudes (vt. tabel 10) võib teha mõningaid järeldusi ka korrelatsiooni puudumise alusel. Nii näiteks on huvitav täheldada, et nimisõnade ja tegusõnade aktiivsus tekstis ei ole omavahel seotud (r = -0,12), samuti puudub oluline seos nimisõnade ja kaassõnade sageduste vahel (r = 0,24). Esimesel juhul tuleb ilmsiks juba varem mainitud tõsiasi, et nominaalsus ja verbaalsus võivad esineda kõrvuti ühes ja samas stiilis, kui neid defineerida sõnaliikide sageduste ja mitte nende suhte alusel. Teisel juhul võib tõdeda, et kaassõnade kasutamine, mis kaasneb nimisõnade kasutamisega, allub individuaalsele valikule ja võib seetõttu olla autoristiile eristavaks teguriks.

Korrelatsioonide põhjalikumaks iseloomustamiseks on võimalik pöörduda faktoranalüüsi poole, mis aitab sõnaliikidevahelisi seoseid esitada kompaktsel kujul. Meie materjali põhjal arvutatud kolm peamist faktorit on järgmised: nominaalsuse faktor (positiivse faktorkoormusega nimisõnal), verbaalsuse faktor (bipolaarne faktor, mille teiseks pooluseks on kvalitatiivsus põhiliste koormustega omadussõnal, määrsõnal ja arvsõnal) ning analüütilisuse faktor (peamiselt kaas- ja sidesõnade aktiivsuse arvel).+

V i i d a t u d k i r j a n d u s

A n t o s c h , F. The Diagnosis of Literary Style with the Verb-Adjective Ratio. - Statistics and Style. Edited by L. Doležel and R.W. Bailey. New York, 1969, 57-68.

B o d e r , D.P. The Adjective-Verb Quotient: A Contribution to the Psychology of Language. - Psychological Record, III, 1940, 309-343.

B u s e m a n n , A. Die Sprache der Jugend als Ausdruck der Entwicklungsrhytmik. Jena, 1925.

B u s e m a n n , A. Stil und Charakter. Untersuchungen zur Psychologie der individuellen Redeform. Meisenheim/Glan, 1948.

E n n i k o , K., M e i m a n , S. Sõnaliikide kvantitatiivne esinemus A.H. Tammsaare romaani "Tõde ja õigus" I köite autorikõnes. TPedI kursusetöö, juh. A. Villup. Tln., 1975.

F i s c h e r , H. Entwicklung und Beurteilung des Stils. - Mathematik und Dichtung. München, 1965, 171-183.

F r i e s , Ch. The Structure of English. New York, 1952.

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

+ Eesti keele sõnaliikide faktoranalüüsi tulemusi tutvustab autorite ettekanne IV rahvusvahelisel fennougristide kongressil Budapestis 1975. (Faktoranalüüsi kohta lähemalt vt. Тулдава, 1976).

K e l e m e n , B. A propos des caractéristiques des styles de la langue à la lumière de la statistique linguistique. - Revue roumaine de linguistique, t. 9, No. 6. Bucarest, 1964, 621-624.

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

K r a u s , J. On the Stylistical-Semantic Analysis of Adjectives in Journalistic Style (A Quantitative Approach). - Prague Studies in Mathematical Linguistics.Vol. 4. Prague, 1972, 95-106.

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

Latviešu valodas biežuma vārdnīca. Atb. red. T. Jakubaite. 3. sējums. Daiļliteratūra. 1. daļa. Rīgā, 1972.

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

L y o n s , J. Towards a 'Notional' Theory of the 'Parts of Speech'. - Journal of Linguistics, 1966, No. 2, 209-236.

M i h k l a , K. jt. Eesti keele lauseõpetuse põhijooned, I. Lihtlause. Toim. E. Nurm. Tln., "Valgus", 1974.

M u l l e r , Ch. Etude de statistique lexicale. Le vocabulaire du théatre de Corneille. Paris, "Larousse", 1967.

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

R e n s k ý , M. The Noun-Verb Quotient in English and Czech. - Philologica Pragensia, VIII. Prague, 1965, 289-302.

Z s i l k a , T. Štylistický rozbor básnickej zbierky štatistickou metódou. - Sborník pedagogickej fakulty v Nitre, 11/1967. Spoločenské vedy (jazyk, literatúra). Bratislava, 1967, 143-162.

Z s i l k a , T. A stílus hirértéke. Bratislava, "Madách Könyvkiadó", 1973.

Z s i l k a , T. Stilisztika és statisztika. Budapest, "Akadémiai Kiadó", 1974.

T š i t e l o v á , M. On the So-Called Vocabulary Richness. - Prague Studies in Mathematical Linguistics. Vol. 3. Prague, 1972, 103-120.

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

T i i t , E. Matemaatilise statistika tabelid, II. Tartu, TRÜ, 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). TPedI võistlustöö, juh. A. Villup. Tln., 1974.

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

T u l d a v a , J. Statistilised testid keeleteaduses. - Linguistica, II. Tartu, 1970, 125-196.

T u l d a v a , J. Sõnavara statistilisest struktuurist. - Linguistica, III. Tartu, 1971, 211-248.

T u l d a v a , J. Korrelatsioonanalüüs keeleteaduses (1). - Linguistica, IV. Tartu, 1972, 163-198.

V a l g e , J. Über die Frequenz der Wortarten in drei funktionalen Stilen. - Generatiivse Grammatika Grupi aastakoosolek. Teesid. Tartu, 1970, 44-45.

V a l g e , J. Poliitika- ja spordialaste ajalehetekstide statistiliste struktuuride erinevusest. - Keel ja Struktuur, 7. Tartu, 1972, 28-38.

V a l g m a , J., R e m m e l , N. Eesti keele grammatika. Käsiraamat. Tln., "Valgus", 1968.

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

V i l l u p , A. Adverbaalse sõnavara statistilisest struktuurist. - Emakeele Seltsi aastaraamat, 19-20. Tln., 1975, 73-90.

А н д р е е в Н.Д. Статистико-комбинированные методы в теоретическом и прикладном языковедении. М.-Л., "Наука", 1967.

А н д р е е в а Л.Д. Статистико-комбинаторные типы словоизменения и разряды слов в русской морфологии. Л., "Наука", 1969.

Б е к т а е в К.Б., Л у к ь я н е н к о в К.Ф. О законе распределения единиц письменной речи. – Статистика речи и автоматический анализ текста. Л., "Наука", 1971, 47–112.

Б е к т а е в К.Б., П и о т р о в с к и й Р.Г. Математические методы в языкознании, ч. 2. Алма-Ата, 1974.

Б е л о н о г о в Г.Г. Распределение частот появления флективных классов русских слов. – Проблемы кибернетики. Вып. II. М., 1964, 189–198.

В и н о г р а д о в В.В. Современный русский язык. М., 1938.

В у к о в и ч Й. К проблеме классификации частей речи.– "Вопросы языкознания", 1972, № 5, 49–61.

Г а л ь п е р и н И.Р. Информативность единиц языка. Пособие по курсу общего языкознания. М.,"Высшая школа", 1974.

Г о л о в и н Б.Н. Язык и статистика. М., "Просвещение", 1971.

Г о л о в и н Б.Н. Опыт применения корреляционного анализа в изучении языка. – Вопросы статистической стилистики. Отв. редакторы Б.Н.Головин и В.И.Перебейнос. Киев, "Наукова думка", 1974, 5–16.

Ж и р м у н с к и й В.М. О природе частей речи и их классификации. – Вопросы теории частей речи. На материале языков различных типов. Л.,"Наука", 1968, 7–32.

З о р е ф М.Г. Машинные основы и машинная морфология в немецко-русском автоматическом словаре. Автореф.канд. дисс. Кишинев, 1972.

К л о ч к о в а Э.А. О распределении классов слов в некоторых функциональных стилях русского языка. – Вопросы славянского языкознания. Саратов, Изд. СГУ, 1968.

К л я в и н я С.П. О нормальности распределения частей речи в текстах (на материале современной латышской публицистики). – Latviešu valodas salīdzināmā analīze. LVU Zinātniskie raksti. 118. sējums. Rīgā, 1969.

Кожина М.Н. О речевой системности научного стиля сравнительно с некоторыми другими стилями. Учебное пособие. Пермь, 1972.

Кривоносов А.Т. Структура языка и система классов слов. - "Вопросы языкознания" 1973, № 4, 86-97.

Лукьяненков К.Ф. Об одном способе автоматического выявления лексической однородности текста. - Лингвостатистика и автоматический анализ текстов (сборник теоретических статей). Минск, Изд. МГПИИЯ, 1974, 325-338.

Лятина А.М. Опыт статистического анализа языка писателя. Автореф. канд. дисс. Л., 1968.

Мистрик Й. Математико-статистические методы в стилистике. - "Вопросы языкознания", 1967, № 3, 42-52.

Пак Х.Я. О некоторых статистико-комбинаторных характеристиках функциональных классов (на материале эстонского языка). - Статистико-комбинаторное моделирование языков. М.-Л., "Наука", 1965, 483-489.

Папина А.Ф., Иванова-Маркова Л. П. Борьба категорий качественности и действенности в поэтической речи первой половины XIX века. - Вопросы статистической стилистики. Киев, "Наукова думка", 1974, 237-243.

Раскина А.А., Чепиго Т.С. Фактографическая ИПС и система микроуниверсалий (на материале русских словоформ). - "Научно-техническая информация. Серия 2. Информационные процессы и системы". М., 1970, №12, 21-28.

Ревзин И.И. Метод моделирования и типология славянских языков. М., "Наука", 1967.

Сиротинина О.Б. Некоторые жанрово-стилистические изменения советской публицистики. - Развитие функциональных стилей современного русского языка. М., "Наука", 1968, 101-125.

Стеблин-Каменский М.И. Спорное в языкознании. Л., Изд. ЛГУ, 1974.

Сумарокова Л.Н. К вопросу о критериях простоты грамматических систем. -Логика и методология науки. М., "Наука", 1967, 86-91.

Суник О.П. Общая теория частей речи. М., "Наука", 1966.

С у н и к О.П. Вопросы общей теории частей речи. – Вопросы теории частей речи. На материале языков различных типов. Л., "Наука", 1968, 33–48.

Т а р л и н с к а я М.Г. Вспомогательный метод установления авторства стихотворного текста. – Проблемы прикладной лингвистики. Тезисы межвузовской конференции. Вып. 2. М., 1969, 300–304.

Т и щ е н к о В. Частота частин мови в різних функціональних стилях сучасної української мови. – Питання структурної лексикології . Киев, 1970.

Т о б и а с Т. Части речи в эстонском языке. – Сообщения по машинному переводу. Вып. 1. Таллин, 1962, 90–96.

Т у л д а в а Ю. Опыт квантитативного анализа художественного стиля. – Studia metrica et poetica, 1. Tartu, 1976 (trükis).

Т у р ы г и н а Л.А. Использование статистических методов при определении изменений в грамматическом строе языка (на материале английских и американских публицистических текстов). – Научная конференция "Проблемы синхронного изучения грамматического строя языка". Тезисы докладов и сообщений. М., 1965, 190–192.

Ы й м Х. О лексических категориях в глубинной структуре. – Проблемы моделирования языка, 3.2. Учен. зап. ТГУ, вып. 228. Тарту, 1969, 197–208.

Ф р у м к и н а Р.М. О законах распределения слов и классов слов. – Структурно-типологические исследования. М., "Наука", 1962.

Ш а й к е в и ч А.Я. Опыт статистического выделения функциональных стилей. – "Вопросы языкознания", 1968, № 1, 64–76.

Ш т е й н ф е л ь д т Э.А. Частотный словарь современного русского литературного языка. Таллин, 1963.

Я к у б а й т и с Т.А. Зависимость вида распределения языковых единиц от величины выборки. – Математические методы в языкознании. Рига, "Звайгзне", 1969, 53–58.

Я к у б а й т и с Т.А. Части речи в художественных текстах. – Известия АН Латвийской ССР, 1974, № 10 (327), 116–126.

Я к у б а й т и с Т.А., С т у р и т е Б.А. О статистической однородности текстов. – Вопросы статистической стилистики. Киев, "Наукова думка", 1974, 43–54.

О ЧАСТОТНОСТИ ЧАСТЕЙ РЕЧИ В АВТОРСКОЙ РЕЧИ ХУДОЖЕСТВЕННОЙ ПРОЗЫ

Ю. Тулдава, А. Виллуп

Р е з ю м е

В статье приводятся данные о распределении частот частей речи в тексте и словаре эстонского языка. Анализу подвергаются 10 выборок по 5000 словоупотреблений (10 х 5000) из авторской речи современной эстонской художественной прозы. В работе рассматриваются общие и индивидуальные тенденции в употреблении отдельных частей речи (по тексту - см.табл.1-2, по словарю - табл. 7) и их группировок, таких как изменяемые/неизменяемые слова (табл. 5) и полнозначные/неполнозначные слова. Тексты сравниваются на основе индексов соотношений частот частей речи (табл. 9) и выявляются корреляции между частотами частей речи в тексте (табл. 10). Результаты сравниваются с соответствующими данными других подъязыков эстонского языка (табл. 3 и 5) и других языков (табл. 4 и 6). На основе сопоставления частот частей речи в тексте и словаре определяются функциональная нагрузка ("индекс итерации", или средняя частота по формуле: t/s, где t - частота в тексте, s - частота в словаре) и информативная нагрузка частей речи в тексте (по формуле p_s/p_t, где p_s - относительная частота части речи в словаре, p_t - относительная частота в тексте). (См. табл. 8). Рассматривается вопрос о репрезентативности материала и статистической достоверности результатов. Статистические данные характеризуются статистическим рядом, арифметической средней, абсолютной и относительной ошибкой, доверительным интервалом и показателем однородности "хи-квадрат" (на уровне достоверности 95 %).

Суммируя основные результаты, приводим данные о частоте появления частей речи в авторской речи современной эстонской художественной прозы на уровне текста и словаря словоформ (данные рассчитаны в среднем на выборку с объемом 5000 словоупотреблений):

Части речи	Относительная частота		Информационная нагрузка
	в тексте p_t (%)	в словаре p_s (%)	p_s/p_t
Имя существительное	3I,7	44,2	I,4
Имя прилагательное	6,0	9,5	I,6
Имя числительное	I,I	I,3	I,I
Местоимение	II,4	4,4	0,4
И м е н а вместе взятые	50,2	59,4	I,2
Глагол:			
инфинитивные формы	8,2	I2,0	I,5
финитивные формы	I4,3	I4,6	I,0
Все формы глагола	22,5	26,6	I,2
И з м е н я е м ы е слова вместе взятые	72,7	86,0	I,2
Наречие	I5,8	I0,7	0,7
Послелог и предлог	3,I	2,2	0,7
Союз	8,2	0,9	0,I
Междометие	0,2	0,2	I,0
Н е и з м е н я е м ы е слова вместе взятые	27,3	I4,0	0,5

STATISTICAL ANALYSIS OF THE PARTS OF SPEECH IN ESTONIAN FICTION

J. Tuldava, A. Villup

S u m m a r y

The article provides data on the frequency of the parts of speech of the Estonian language in text and vocabulary. 10 samples of 5,000 words each of Modern Estonian prose fiction (non-conversational material) are analyzed. General and individual tendencies in the use of the different parts of speech (on the text level - see Tables 1-2, on the vocabulary level - Table 7) and groups of them, such as inflected/uninflected words (Table 5) and "full-meaning" and form words, are examined. The texts are compared on the basis of the ratios of the parts of speech (Table 9) and the correlations between the parts of speech in a text are fixed (Table 10). The results are compared with corresponding data of other sublanguages of Estonian (Tables 3 and 5) and those of other languages as well (Tables 4 and 6). On the basis of contrasting the parts of speech in the text and in the vocabulary, their functional load (according to the formula t/s, where t - frequency in the text, s - frequency in the vocabulary) and informational load in the text (p_s/p_t - where p_s - relative frequency in the vocabulary and p_t - relative frequency in the text) are calculated (Table 8). The problem of the representativeness of the material and statistical significance of the results is treated. Statistical data are characterized by series of frequencies, arithmetical mean, standard error and confidence intervals on the 95 % level. Homogeneity is verified by the Chi-Square-Test.

In the following table data on the frequency of use of the parts of speech in Modern Estonian fiction (non-conversational material) are presented.

Frequency of Use of Forms in Text and Vocabulary

Parts of speech	Relative frequency		Informational load
	in text p_t (%)	in vocabulary p_s (%)	(p_s/p_t)
Nouns	31.7	44.2	1.4
Adjectives	6.0	9.5	1.6
Numerals	1.1	1.3	1.1
Pronouns	11.4	4.4	0.4
Totals	50.2	59.4	1.2
Verbs:			
non-infinite forms	8.2	12.0	1.5
finite forms	14.3	14.6	1.0
Totals	22.5	26.6	1.2
All inflected forms	72.7	86.0	1.2
Adverbs	15.8	10.7	0.7
Postpositions and prepositions	3.1	2.2	0.7
Conjunctions	8.2	0.9	0.1
Interjections	0.2	0.2	1.0
All uninflected forms	27.3	14.0	0.5

EESTI KEELE ILUKIRJANDUSPROOSA AUTORIKÕNE SÕNAVORMIDE SAGEDUS-SÕNASTIK

Ü. Kaasik, J. Tuldava, A. Villup, K. Ääremaa

Käesoleval ajal on koostatud sagedussõnastikud peaaegu kõigi kultuurkeelte, sealhulgas ka meie naaberkeelte - vene, läti, soome ja rootsi keele kohta. Eesti keele alalt on ilmunud mõned kirjanikusõnastikud sagedusandmetega (Piir, 1963; Tauli, 1964; Vihma, 1970), kuid suurema ulatusega sagedussõnastik seni puudus. Alles 1972. a. asuti esimese eesti keele sagedussõnastiku kavandamisele. Algatus tuli Tartu Riikliku Ülikooli keelestatistikarühma liikmetelt ja Tallinna Polütehnilise Instituudi arvutuskeskuse teaduslikult juhendajalt L. Võhandult, kelle eestvedamisel sai teoks tekstide perforeerimine. Eeltööd kestsid 1975. aasta alguseni. Samal aastal koostati TRÜ arvutuskeskuses programm (autor Ü. Kaasik), mis realiseeriti Ü. Kaasiku ja K. Ääremaa juhendamisel. Käesolevas kogumikus avaldatakse sagedussõnastiku esimene osa.

Esialgsete plaanide kohaselt kavatsetakse koostada eraldi nelja allkeele sagedussõnastikud ja seejärel "üldkeele" sõnastik, s. o. nelja allkeele koondsõnastik. Sagedussõnastik peaks hõlmama järgmised allkeeled: ilukirjandusproosa autorikõne, ilukirjandusproosa tegelaskõne ning draamadialoogi, ajalehekeele ja teaduskeele. Ilukirjandusproosa autorikõne ja tegelaskõne eristamise vajadus on tingitud nende allkeelte sõnavara erinevast laadist. Igast allkeelest võetakse ca 100 000-sõnaline valim (väljavõtukogum, väljavõte), kusjuures koguvalim koosneb parema representatiivsuse tagamiseks paljudest osavalimitest.

Esimese allkeelena on käesolevaks ajaks statistiliselt

tõõdeldud eesti tänapäeva ilukirjandusproosa **a u t o r i k õ n e** . Valim koosneb 20 osavalimist à 5000 sõnet kahekümnest eri teosest, mis on ilmunud pärast 1960. a. Teoste valiku osas (1972. a.) abistasid koostajaid kirjandusteadlased H. Puhvel ja A. Vinkel, kellele siinkohal avaldatakse siirast tänu. Ilukirjandusproosa autorikõne sagedussõnastiku materjal pärineb järgmistest teostest.

I. Romaanid

1. A. Beekman, Kartulikuljused (1968)
2. V. Gross, Pinginaabrid (1965)
3. A. Hint, Tuuline rand IV (1966)
4. H. Kiik, Tondiöömaja (1970)
5. J. Kross, Kolme katku vahel I (1970)
6. P. Kuusberg, Südasuvel (1966)
7. L. Promet, Primavera (1971)
8. V. Saar, Ukuaru (1969)
9. H. Sergo, Põgenike laev (1966)
10. R. Sirge, Kolmekesi lauas (1970)
11. M. Traat, Tants aurukatla ümber (1971)
12. E. Vetemaa, Väike romaaniraamat (1968)

II. Novellid, lühijutud, miniatuurid

13. A. Kaal, Saaremaa laastud II (1970)
14. T. Kallas, Puiesteede kummaline valgus (1968)
15. J. Peegel, Lühikesed lood (1970)
16. J. Tuulik, Vana loss, Abruka lood (1972)
17. A. Valton, Luikede soo, Karussell (1971)
18. M. Unt, Kuu nagu kustuv päike (1968)

III. Reisikirjad

19. J. Kross, E. Niit, Muld ja marmor (1968)
20. J. Smuul, Jaapani meri, detsember (1965)

Nagu varem nimetasime, võeti igast teosest 5000-sõneline valim, mis omakorda jaguneb viieks juhuslikuks valitud tekstilõiguks à 1000 sõnet teose erinevatest osadest. Valimisse lülitati ainult autorikõne. Materjal analüüsiti grammatiliselt ja indekseeriti homonüümid (A. Villupi poolt).

Seejärel perforeeriti materjal viierajalisele lindile. Edasi toimus töötlemine juba automaatselt: perforeeritud materjali põhjal koostas arvuti sõnavormide sagedussõnastikud iga autori kohta eraldi (osasõnastikud) ja kõigi autorite koondsõnastiku. Osasõnastikud koostati nii alfabeetilises kui ka sagedusjärjestuses. Koondsõnastiku väljastas arvuti kolmes variandis: alfabeetilises järjestuses, sagedusjärjestuses absoluutse esinemissageduse (F) järgi ja sagedusjärjestuses "modifitseeritud sageduse" ehk kasutuskoefitsiendi (U) järgi. Kõik sõnastikud on koostatud sõnavormide tasandil. Sõnavormide koondamine põhivormide alla (nn. lemmatiseerimine) toimub käsitsi. Sel teel saadav lekseemide sagedussõnastik avaldatakse kogumiku järgmises väljaandes.

Pärast töötlust arvutil selgus, et valimi täpne üldmaht oli 99 898 sõnet. Sellest loendati 30 733 eri sõnavormi. Üks kord esinevaid sõnavorme (nn. hapax legomena) oli 21 760, mis moodustab 70,8 % kogu sõnavormide sõnastikust ja 21,8 % tekstist. Kaks korda esinevaid sõnavorme oli 3885 (12,6 % sõnastikust, 7,8 % tekstist) ja kolm korda esinevaid sõnavorme 1628 (5,3 % sõnastikust, 4,9 % tekstist). Sagedusjärjestuses on nii osasõnastikes kui ka koondsõnastikus esikohal sidesõna ja üldsagedusega 3221 (3,22 % tekstist). Järgnevad ta (nimetav kääne) 1602, on 1429, ei 1329, et 1203, oli 1116, kui 989, see 711, oma 592, nagu 536. Nimetatud sõnavormid esinesid kõigis 20 osavalimis, kuid mõnevõrra erinevate absoluutsagedustega osavalimite lõikes.

Keelestatistikas kujunenud tavade kohaselt arvestatakse sagedussõnastike koostamisel esinemissageduse s t a b i i l s u s t eriliste stabiilsuskordajate alusel, võttes arvesse esinemuse kõigis osavalimites ja hajuvuse keskmise ümber. Kõige tuntum stabiilsuskordaja on prantsuse keelestatistiku A. Juilland'i koefitsient D (1964), mis arvutatakse järgmiselt:

$$D = 1 - \frac{v}{\sqrt{k-1}},$$

kus v tähistab variatsioonikordajat ja k osavalimite arvu (tingimusel, et osavalimite mahud on võrdsed). Koefitsient võib omandada väärtusi 0 ja 1 vahel (mõnel juhul esineb ka negatiivne väärtus, mis loetakse nulliks). Seda koefitsienti võib vaadelda iseseisva näitajana, mida kasu-

tatakse näit. keelte õpetamiseks mõeldud miinimumsõnastike koostamisel. Tavaliselt esitatakse stabiilsusnäitaja nn. kasutuskordaja U (inglise sõnast usage) vahendusel:

$$U = D \cdot F ,$$

kus F tähistab antud sõna sagedust üldvalimis. Sel juhul saadakse stabiilsuskoefitsiendi abil "modifitseeritud" sagedus, mis paremini kui registreeritud üldsagedus (F) väljendab antud sõna tegelikku suhtelist sagedust, võrreldes teiste sõnadega. See on eriti oluline väikeste ja keskmiste suurusega sagedussõnastike koostamisel, kui on tegemist suhteliselt väikese üldvalimiga. Peab silmas pidama, et U väärtus on reeglina väiksem F väärtusest (U = F ainult sel juhul, kui D = 1, s. o. antud sõna esineb maksimaalse stabiilsusega - täpselt ühepalju kõigis osavalimites). Näiteks sõnavorm on, mille üldsagedus on 1429, jaotub osavalimites järgmiselt: 23-54-21-129-21-86-29-76-15-28-163 - 114-76-41-119-50-79-71-144-90. Stabiilsuskordaja D väärtuseks saame 0,828, mis näitab, et antud sõnavorm ei esine eriti stabiilselt. Modifitseeritud sagedus U = 0,828.1429= = 1183. Võrdluseks võib tuua sõnavormi ei üldsagedusega 1329. See sõnavorm esineb stabiilsemalt: D = 0,950 ja U = 0,950· ·1329 = 1262. Järelikult viib U väärtus sõnavormi ei sõnavormist on sagedusjärjestuses ettepoole ja seda võib suure tõenäosusega oodata juhul, kui suurendaksime oluliselt oma valimi üldmahtu. (Sõnavormi on sageduste suurem kõikumine osavalimite lõikes on tingitud selle konkureerimisest vormiga oli.) Seepärast ongi põhimõtteliselt õigem asetada sõnavormid sagedussõnastikus sagedusjärjekorda kasutuskordaja U väärtuste alusel (nende väärtuste kahanevas reas). Niiviisi on koostatud viimasel ajal paljud sagedussõnastikud, näit. prantsuse, hispaania, rumeenia, slovaki ja rootsi keele sõnastik (Juilland jt., 1964; 1965; 1970; Mistrík, 1969; Allén, 1970). Stabiilsus- ja kasutuskordajaid arvestatakse ka uuemates keelestatistilistes uurimustes vene ja ukraina keele kohta (vt. näit. Алексеев, 1969; Дарчук, 1974). Käesoleva sõnastiku koostamise käigus arvutati samuti D ja U väärtused ning üks sagedussõnastiku variantidest moodustatigi U alusel. Et see sõnastikuvariant meie suhteliselt väikese valimi puhul on kahtlemata usal-

datavam kui lihtsalt esinemissageduse (F) alusel järjestatud sõnastik, siis esitatakse see alljärgnevalt ilukirjandusproosa autorikõne sagedussõnastiku esimese osana, lähtudes sõnavormide tasandist. Sõnastiku mahtu piirati sellega, et sõnastikku lülitati ainult need sõnavormid, mille modifitseeritud sagedus on üle 1 (U > 1). Selliseid sõnavorme on kokku 2927. Sõnavormid on järjestatud kasutuskordaja U põhjal, kusjuures järjekorranumber (astak) märgitakse iga kümnenda sõnavormi ette.

Sõnastikus leiavad kasutamist järgmised lühendid:

(s) - substantiiv
(a) - adjektiiv
(n) - numeraal
(p) - pronoomen
(v) - verb
(s^x), (a^x), (p^x), (v^x) - substantiiv, adjektiiv, pronoomen või verbi käändeline vorm väljendverbi komponendina
(d_1) - iseseisev adverb
(s/d_1) - adverbiaalses funktsioonis kasutatud nimisõnavorm (pooladverb)
(d_2) - modaaladverb (rõhumäärsõna)
(d_3) - prefiksaaladverb (abimäärsõna)
(pp) - pre- või postpositsioon
(k) - konjunktsioon
(k^x) - ühendkonjunktsiooni komponent
(i) - interjektsioon
(nom) - nominatiiv
(gen) - genitiiv
(part) - partitiiv
(ill) - illatiiv
(in) - inessiiv
(el) - elatiiv
(all) - allatiiv
(ad) - adessiiv
(sg) - singular
(pl) - pluural
(neg) - eitava kõne vorm
(imp) - imperatiiv
(atr) - (mineviku) kesksõna atributiivses või predikatiivses funktsioonis
(t) - määruslik tarind (lauselühend)

Lühendeid kasutatakse leksikaalsete ja grammatiliste homonüümide eristamiseks, samuti muudel juhtudel, kus see on vajalik sõnavormi või funktsiooni määramiseks. Sealjuures peetakse silmas järgmist.

Käändsõnavormidele lisatakse vajaduse korral kas sõnaliigi või käändenimetuse lühend, näit. viis (s), viis (n), rahvast (part), rahvast (el). Tavalistest nominaalvormidest lahus esitatakse väljendverbide+ komponendid koos sõnaliigi või vormi lühendiga, näit. appi (s^X), osa ($part^X$), otsa (ill^X). Ühtelangevate arv- ja asesõnavormide puhul varustatakse sõnaliigi lühendiga viimased, näit. teine (p), ühel (p), arvsõnad aga ainult erandjuhtudel, nagu pool (n) (et hoida viimast lahus homonüümsest nimi- ja kaassõnast).

da-infinitiivi ja imperatiivi ainsuse 2. pöörde homonüümsuse korral tähistatakse vastava lühendiga imperatiivivorm kui harvemini kasutatav, näit. seada (imp) verbist saama. Mineviku kesksõnadest märgitakse ära ainult need, mis esinesid tekstis määrusliku tarindina (t), täiendina või öeldistäitena (atr). Ülejäänud juhtudel esines kesksõna öeldise funktsioonis (liitaja osisena või ka üksi).

Omadussõna ablatiivivormiga homonüümsed lt-liitelised iseseisvad adverbid esitatakse üldiselt ilma sõnaliigi lühendita. Kui vaja, lisatakse see samakujulisele adjektiivile. Erandiks on juhud, kus üht ja sama adverbi kasutatakse nii iseseisva kui ka rõhumäärsõnana, näit. lihtsalt (d_1) ja lihtsalt (d_2). Öeldu kehtib samuti teiste määrsõnatüüpide suhtes (peale lt-liiteliste), näit. otse (d_1) ja otse (d_2). Ainult abimäärsõnad, mis sageli on homonüümsed kaassõnadega, tähistatakse alati vastava lühendiga (d_3). Samuti toimitakse kaassõnade (pp) puhul. Pre- ja postpositsioone pole käesoleval juhul eristatud, seda tehakse lemmatiseeritud sõnastikus.

Sidesõnadest varustatakse lühendiga (k) need, mis on homonüümsed kaassõnade või rõhumäärsõnadega (harvemini teiste sõnaliikide sõnadega), samuti ühendsidesõna komponendid (k^X), mis üksikult konjunktsioonina ei esine, näit. ei, ka, mitte, nii jt. (ühendites, nagu ei ... ega, nii...

+ Vt. Rätsep, 1973.

kui ka, mitte ainult ... vaid ka). Kui sidesõna funktsioneerib niihästi iseseisvalt kui ka sidesõnaühendi osisena, esitatakse sõnastikus sõna sagedus summeeritult (s. t. k ja k^{x} näitajad koos).

Leksikaalsetele homonüümidele lisatakse mõnel juhul sulgudes sõna tähendus, seletus või ka üks põhivormidest, näit. aru (mõistus), saar (veekogus), tee (liikl.), peos (pihk), lõi (lööma). Kui aga kahest või enamast homonüümist esineb sõnastikus ainult kõige levinum, ei peeta vajalikuks tähendust märkida, näit. liiv, mees, kohad (sõnast koht), talle (=temale).

SÕNAVORMIDE SAGEDUSSÕNASTIK

(järjestus kasutuskordaja U järgi)

	U			U	
1.	3138	ja (k)		194	polnud
	1491	ta (nom)		187	midagi
	1297	ei		187	ära (d_3)
	1187	on		185	teda
	1186	et		182	tema (gen)
	951	kui		176	oleks
	926	oli		168	kas
	669	see		168	välja (d_3)
	566	oma		165	üle (pp)
10.	515	nagu	40.	164	mitte
	509	mis		156	kuid (k)
	407	siis		155	olid (pl)
	387	ka		154	ta (gen)
	359	aga (d_2)		154	tal
	354	seda		153	küll (d_2)
	315	ning		153	pärast (pp)
	313	aga (k)		152	enam
	308	nii		148	kus
	305	ma		144	mida
20.	292	või (k)	50.	143	talle
	279	kes		140	ise
	276	veel		133	neid
	234	kõik		131	kuidas
	228	nad		130	need
	227	selle		127	nende
	225	ja (d_2)		126	sest (k)
	225	juba		126	siin
	205	pole		124	seal
	204	nüüd		123	kõige (d_1)
30.	196	ainult	60.	120	mees

	117	tuli (v)		66	teha
	114	ega (k)		66	võib-olla
	113	isegi (d_2)		65	juurde (pp)
	112	vastu (pp)		65	siiski
	108	tema (nom)		64	minu
	106	end		64	praegu
	102	tagasi (d_3)		64	üles (d_3)
	101	all (pp)		63	alati
	101	ju		63	mille
70.	99	kogu (p)	110.	62	ikka
	98	poole (pp)		62	sai (v)
	96	tuleb		61	hea
	95	palju (d_1)		61	vaatas
	94	läks		60	alles (d_1)
	94	sellest		60	läbi (pp)
	93	üks		59	ees (pp)
	92	kuid (d_2)		58	ajal
	91	me (nom)		58	olnud
	88	jälle		58	ometi
80.	83	just	120.	58	selles
	82	meie (gen)		58	vahel (pp)
	81	väga		58	ümber (pp)
	80	hakkas		57	sa
	80	peale (pp)		56	kinni (d_3)
	77	mulle		56	kokku (d_3)
	76	mu		56	läbi (d_3)
	73	jäi		55	nii (k^x)
	72	keegi		55	siis (d_2)
	72	muidugi		55	ütles
90.	72	vaid (k)	130.	54	eest (pp)
	72	võib		54	peaaegu
	70	olla		54	poleks
	69	juures (pp)		54	saab
	69	peab		54	uuesti
	69	rohkem		53	aega (part)
	68	kaks		53	iga (p)
	68	maha (d_3)		53	kunagi
	68	mind		53	mööda (pp)
	68	vastu (d_3)		53	tegi
100.	68	vist	140.	52	edasi (d_3)

52 kuigi (k)
52 mingi
52 mul
52 pisut
52 tąga (pp)
51 kohe
51 silmad (s)
50 ole (neg)
50 sel
150. 49 mina
48 naine
48 suur
47 ennast
47 võis (v)
46 võttis
45 lahti (d_3)
45 olin
45 sisse (d_3)
45 suure
160. 44 istus
44 küllap
44 läheb
44 mõni
44 peale (d_3)
44 õhtul
43 alla (pp)
43 ikka (d_2)
43 näha
43 ühe
170. 42 asi
42 kelle
42 minna
42 sinna
41 endale
41 ette (d_3)
41 jah
41 vana
40 isa (nom)
40 miks
180. 40 minema
40 poiss
40 saa (neg)
40 üldse
39 enda
39 kolm
39 kord (d_1)
39 kuhu
39 käis (v)
39 pidi (v)
190. 39 saanud (neg)
39 üle (d_3)
38 jääb
38 poolt (pp)
38 päris (d_1)
37 enne (pp)
37 kohal (pp)
37 koos (d_1)
37 korda (part)
37 mehe
200. 37 rääkis
37 tea (neg)
37 tundis
36 kui (d_1)
36 neile
36 tõesti
35 aastat
35 ehk (d_2)
35 inimene
35 jaoks (pp)
210. 35 liiga (d_1)
35 mehed
35 mööda (d_3)
35 olema
35 pani
35 saanud
35 täis (a)
34 aeg
34 aru (mõistus;part^x)
34 hakkab
220. 34 hommikul

	34	kaasa (d_3)
	34	kõrval
	34	päeval
	34	seisis
	34	vaadata
	33	nägu
	33	päike
	33	teada
	33	teeb
230.	33	täiesti
	32	hoopis
	32	hästi
	32	järgi (pp)
	32	kohta (pp)
	32	mõnikord
	32	neist
	32	panna
	32	siia
	32	võtta
240.	31	hiljem
	31	mõne
	31	mõned
	31	otse (d_2)
	31	saada
	31	sealt
	31	vahel (d_1)
	31	vaid (d_2)
	30	ema
	30	hoopis (d_2)
250.	30	ikkagi
	30	inimesed
	30	kust
	30	käed
	30	läinud
	30	meelde (s^x)
	30	olnud (neg)
	30	otsekui
	30	ringi (d_3)
	30	ukse
260.	29	jäänud
	29	kord (s)
	29	meie
	29	teine
	29	või (d_2)
	28	andis
	28	elu (gen)
	28	koju
	28	lihtsalt (d_2)
	28	neil
270.	28	olen
	28	taha (pp)
	28	tuul
	28	täna (d_1)
	28	vaja (d_3)
	27	ent (k)
	27	inimeste
	27	meid
	27	naise
	27	olevat
280.	27	peal (pp)
	27	sellepärast
	27	teab
	27	teine (p)
	27	tulid (pl)
	27	varsti
	27	veidi
	27	õieti (d_2)
	26	alla (d_3)
	26	eriti
290.	26	ette (pp)
	26	inimese
	26	kõike
	26	maa (gen)
	26	millest
	26	muud (part)
	26	näis
	26	peaks (v)
	26	raske
	26	tundus
300.	26	võiks (v)

	26	äkki
	25	igal
	25	nägi
	25	tahtis
	25	tõepoolest
	25	üsna
	24	ees (d_1)
	24	meri
	24	ning (d_2)
310.	24	püsti (d_1)
	24	selja
	24	teadis
	24	uue
	23	enne (d_1)
	23	ent (d_2)
	23	et (d_2)
	23	jõudis
	23	keda
	23	käe
320.	23	muutus (v)
	23	paistis
	23	pea (gen)
	23	rääkida
	23	siit
	23	teadnud (neg)
	23	tunne (s)
	23	võtab
	23	ääres (pp)
	23	üks (p)
330.	23	ümber (d_3)
	22	ajal (pp)
	22	ilus (a)
	22	inimesi
	22	järele (pp)
	22	järele (d_3)
	22	kuidagi (d_2)
	22	kuni (k)
	22	lõpuks (s)
	22	meil
340.	22	mitu
	22	mõtles
	22	selleks
	22	sellele
	22	tahtnud (neg)
	22	vähemalt (d_2)
	22	väike
	21	alt (pp)
	21	astus
	21	kaua
350.	21	läksid (pl)
	21	mingit
	21	mõelda
	21	oleksid (pl)
	21	otsa (ill^x)
	21	parem (a)
	21	tagasi (pp)
	21	tulla
	21	öelda
	20	anda
360.	20	elu
	20	imelik
	20	juhtus
	20	kahe
	20	kodus
	20	käib
	20	meest (part)
	20	meile
	20	natuke
	20	niisama
370.	20	näinud
	20	sõnu
	20	teinud
	20	teisiti
	20	teiste (p)
	20	temaga
	20	too (p)
	20	tulnud
	20	ühel (p)
	19	akna
380.	19	ilma (pp)

19 koguni (d_2)
19 kuskil
19 kõrvale (d_3)
19 laiali
19 laua
19 nägu (part)
19 oleme
19 poeg
19 poisid
390. 19 päev
19 selline
19 tööd (part)
19 uus
19 vahele (pp)
19 õige (a)
18 asju
18 esimese
18 hetkeks
18 igatahes
400. 18 koos (pp)
18 käes (s)
18 maailma (gen)
18 nagu (d_2)
18 niisugune
18 niisugust
18 niiviisi
18 nõnda
18 pea (s)
18 peremees
410. 18 seisab
18 seisma
18 suured
18 tagant (pp)
18 temast
18 tundsin
18 tõi
18 tõmbas
18 tõstis
18 täis (a^x)
420. 18 töö (gen)
18 vaatab
18 vaatama
18 vahelt (pp)
17 aina
17 hakata
17 hakkasid (pl)
17 hääl
17 jutt (jutu)
17 järel (pp)
430. 17 kell
17 käega
17 küllalt
17 maa
17 mõtlesin
17 naised
17 oodata
17 pool (pp)
17 pärast (d_1)
17 samas (d_1)
440. 17 silmi (part)
17 vaevalt (d_2)
17 vähe
17 üksnes
16 aja (gen)
16 jooksis
16 kaugel (d_1)
16 kätte (s^x)
16 jäid (pl)
16 küsis
450. 16 lõpuks (d_1)
16 maja (gen)
16 mere
16 mõte
16 nii (d_2)
16 näeb
16 nägin
16 paar (n)
16 pealegi (d_2)
16 pikk
460. 16 päeva (gen)

16 sain
16 samal
16 samuti (d_2)
16 sellega
16 taha (neg)
16 tõusis
16 tükk
16 umbes
15 ajas (v)
470. 15 annab
15 arvata
15 ei (k^x)
15 esimest
15 hetkel
15 istub
15 jooksul (pp)
15 jäädа
15 kindel
15 külla (gen)
480. 15 nemad
15 oska (neg)
15 pidanud
15 tahab
15 toas
15 tulnud (neg)
15 tunneb
15 tüdruk
15 vana (gen)
15 viis (v)
490. 14 aasta (gen)
14 aastaid
14 aeg-ajalt
14 ammu (d_1)
14 asemel (pp)
14 ega (d_2)
14 hakkavad
14 isa (gen)
14 jalg
14 juttu (part)
500. 14 jõudnud
14 järsku (d_1)
14 ka (k^x)
14 kaugele (d_1)
14 lapsed
14 lausa
14 linna (gen)
14 millega
14 oleksin
14 oligi
510. 14 pika
14 pilgu (s)
14 pilk
14 suutnud (neg)
14 sõita
14 sõna (part)
14 tee (liikl.)
14 tõuseb
14 vend
14 üksi (d_1)
520. 14 ükski (p)
14 üles
14 ütleb
13 aastal
13 asjad
13 ema (gen)
13 enne (k^x)
13 esimesel
13 hall (a)
13 harva (d_1)
530. 13 hoolimata (pp)
13 hulk
13 ilmus
13 inimest
13 jalg
13 kadus
13 kahju
13 kedagi
13 keset (pp)
13 kiiresti
540. 13 kuidagi (d_1)

13 kõige (p)
13 kõrvale (pp)
13 käest (s)
13 käia (v)
13 lugu
13 lähevad
13 maad (part)
13 mehi
13 minust
550. 13 mõtted
13 noor
13 otse (d_1)
13 päeva (part)
13 püüab
13 püüdis
13 rõõmu (part)
13 seejärel
13 seisid (pl)
13 seljas (s/d_1)
560. 13 sellel
13 sulle
13 teised (p)
13 tunda
13 tuppa
13 tõsiselt
13 tõttu (pp)
13 vett
13 välja (d_1)
13 ühel
570. 12 ajada
12 algas
12 endast
12 enese
12 hobuse
12 häält
12 ilmselt
12 istuda
12 jalad
12 juuksed
580. 12 juurde (d_3)
12 kaob
12 kas (k^X)
12 kellel
12 koht
12 korraks (d_1)
12 kätt
12 leiab
12 leida
12 leidis
590. 12 läksin
12 löi (lööma)
12 näoga
12 olgu
12 osanud (neg)
12 paistab
12 paljud
12 paremini
12 pealt (pp)
12 pidas
600. 12 raha (part)
12 seepärast
12 seesama
12 sest (d_2)
12 silmadega
12 sugugi (d_2)
12 taas
12 tee (liikl.;gen)
12 teed (liikl.;part)
12 tegema
610. 12 teise (p, gen)
12 tekkis
12 tuhat
12 tulevad
12 uut
12 vaatasid (pl)
12 viimane
12 võinud
12 vähem (d_1)
12 ühest
620. 12 ühtki (p)

12 üleni
11 all (d_1)
11 arvas
11 elas
11 esialgu
11 esimene
11 hobused'
11 homme
11 juba (d_2)
630. 11 juhtunud
11 kostis (v)
11 kuna
11 kuulnud
11 kõigi
11 laps
11 laual
11 metsa (gen)
11 metsas
11 milleks (d_1)
640. 11 muidu (d_1)
11 neli
11 nimi
11 nähtavasti
11 näo
11 olime
11 ootamatult
11 paar (s)
11 pannud
11 pilt
650. 11 poolest (pp)
11 pähe
11 saaks
11 selge
11 silmaga
11 sisse (pp)
11 suurem
11 suurt
11 sõda
11 tahtnud
660. 11 tahtsin
11 talvel
11 tarvis (d_3)
11 toob
11 tore
11 täpselt
11 vaatasin
11 vahepeal
11 vahetevahel
11 vaikselt
670. 11 varem
11 veelgi (d_1)
11 viimasel
11 öösel
11 üksteise
10 ainult (k^x)
10 aknast
10 algul
10 alla (d_1)
10 asja (part)
680. 10 eemal
10 ehkki
10 elab
10 elada
10 elu (part)
10 justkui
10 jätta
10 kellele
10 kusagil
10 kuulda
690. 10 kõigil
10 külge (pp)
10 laskis
10 lugeda
10 lööb
10 majas
10 meeste
10 miski
10 muutunud
10 mõlemad
700. 10 mõtet

	10	naist
	10	nimelt (d_2)
	10	oled
	10	olevat (t)
	10	ongi
	10	ootas
	10	otsis
	10	paari (n, gen)
	10	peaga
710.	10	peavad
	10	punane
	10	said (pl)
	10	sees (d_1)
	10	suvel
	10	süda
	10	te
	10	teevad
	10	tee (neg)
	10	tegemist
720.	10	teise (gen)
	10	temal
	10	terve (p)
	10	terve (p, gen)
	10	tookord
	10	tundnud (neg)
	10	töö
	10	vaikus
	10	valmis (d_3)
	10	vankri
730.	10	viimase
	10	väikese
	10	õige (gen)
	10	ööl
	9	appi (s^X)
	9	arvatavasti
	9	asus
	9	eesti (a)
	9	eile
	9	elus (s)
740.	9	igaüks
	9	isegi (p)
	9	istusid (pl)
	9	juttu (part^X)
	9	jõudu (part)
	9	jättis
	9	jäänud (atr)
	9	kaudu (pp)
	9	kevadel
	9	kindlasti (d_2)
750.	9	korraga (d_1)
	9	kuhugi
	9	käinud
	9	lauale
	9	leidub
	9	meist
	9	millel
	9	minnes
	9	mitte (k^X)
	9	muutub
760.	9	mõtleb
	9	mõtteid
	9	mõttes
	9	märkas
	9	niisuguse
	9	nägku
	9	oh
	9	omavahel
	9	ootab
	9	otsima
770.	9	palus (v)
	9	peas (s/d_1)
	9	pidada
	9	rahvast (part)
	9	räägiti
	9	samasugune
	9	seegi (p)
	9	sind
	9	sisse (d_1)
	9	süüa
780.	9	süüdi (d_1)

9 taeva
9 taga (d_3)
9 tegelikult
9 tihti
9 tol
9 tugev
9 tõsi
9 viimaks (d_1)
9 võivad
790. 9 võtma
9 võtsin
9 õde
9 õõ
8 ajab
8 Anna
8 anna (neg)
8 antud
8 Eesti (gen)
8 enam-vähem
800. 8 hakka (neg)
8 halli (a, gen)
8 hobune
8 hoida
8 häbi
8 iialgi
8 juhtub
8 juurest (pp)
8 jõudsin
8 järgmisel
810. 8 jäävad
8 kellelegi
8 koha (gen)
8 kohta (part)
8 koolis
8 korda (ill^x)
8 korralikult
8 kõigile
8 kõrge (gen)
8 kõrvuti
820. 8 laseb
8 luges
8 lumi
8 lõpp
8 lõppes
8 lähemale (d_1)
8 mehele
8 metsa (ill)
8 mil (p)
8 millal
830. 8 milles
8 missugune
8 mõjus (v)
8 noored
8 nägema
8 osa ($part^x$)
8 oskab
8 otsa (pp)
8 otsas (pp)
8 paks
840. 8 paneb
8 perenaine
8 pidanud (neg)
8 pidid (pl)
8 polegi
8 päikese
8 räägib
8 rääkisid (pl)
8 saksa (a)
8 sammu (gen)
850. 8 sedagi
8 seletada
8 selga (s/d_1)
8 selgeks (a^x)
8 selgesti
8 selgus (v)
8 sellise
8 silma (ill^x)
8 sina (p)
8 soe (a)
860. 8 su

8 suhu
8 sul
8 surus
8 suuda (neg)
8 suures
8 suurte
8 sõnad (s)
8 taevas (nom)
8 tean
870. 8 teel (liikl.)
8 teineteise
8 tulema
8 tulles
8 tuule
8 uskuda
8 vaadates
8 vaatavad
8 vajus
8 valmis (a)
880. 8 veel (d_2)
8 vesi
8 vihm
8 võimalik
8 võtnud
8 õhtuti
8 õiget
8 ükskõik
8 üksteisele
8 ütelda
890. 8 ütlema
7 aasta
7 arvab
7 arvasin
7 edasi (d_1)
7 endas
7 haaras (v)
7 hakanud (neg)
7 halb
7 harjunud
900. 7 hea (gen)
7 heitis
7 ilm
7 inimestele
7 jala (s)
7 jookseb
7 Juhan
7 järjest (d_1)
7 kadunud
7 kandis (v)
910. 7 kasvab
7 kasvas
7 keeles
7 kella (gen)
7 kellega
7 kellegi
7 kergelt
7 kipub
7 kiriku
7 kombel (s)
920. 7 kuigi (d_1)
7 kukkus
7 kuulis (v)
7 kõigest (d_2)
7 käes (pp)
7 kätega
7 kätte (s)
7 kätte (pp)
7 küla
7 külje
930. 7 lihtne
7 loom
7 majad
7 meetrit
7 mine
7 muidu (d_2)
7 mulje (gen)
7 mõnel
7 märganud (neg)
7 noh
940. 7 nurgas

7 nähtavale (v^X)
7 näiteks (d_2)
7 omal
7 osa
7 otsustas
7 parajasti (d_2)
7 pead (part)
7 pigem
7 pikka (part)
950. 7 pilku (part)
7 pime (a)
7 poja
7 punase
7 puude (pl)
7 päris (d_2)
7 rääkima
7 saata
7 saavad
7 sageli
960. 7 sealsamas
7 seejuures
7 silma (part)
7 silmade
7 silmadesse
7 suu
7 sõidavad
7 sõna
7 südame
7 Tallinna (gen)
970. 7 teineteist
7 teinud (neg)
7 teisi
7 teist (part)
7 teist (el)
7 tekib
7 toa
7 tuba
7 tule (neg)
7 tundnud
980. 7 tänaval
7 uusi
7 vaikne
7 valgus (s)
7 vastata
7 võõras (a, nom)
7 väljas (d_1)
7 õige (d_1)
7 ära (v)
7 ühe (p, gen)
990. 7 üksinda
7 ükskord
7 ütlesin
6 aeglaselt
6 ah
6 ainus
6 arvates
6 asja ($part^X$)
6 astub
6 astuda
1000. 6 astuma
6 avas (v)
6 eks
6 enamasti
6 enesest
6 ettepoole
6 haarab
6 hakanud
6 hakkasin
6 halvasti
1010. 6 hetk
6 hing
6 hoidis
6 ilusa
6 istudes
6 istuvad
6 jalgu (part)
6 jumal
6 jõe
6 kaela (part)
1020. 6 kallal (pp)

6 kauaks
6 kauem
6 keelt (part)
6 keha
6 kena
6 kerge
6 kippus
6 kohaselt (pp)
6 kuni (pp)
1030. 6 kuulsin
6 kõigepealt
6 kõiki
6 kõvasti
6 käisid (pl)
6 käsi (pl)
6 käte
6 kümme
6 küsimus
6 lai
1040. 6 lase (neg)
6 liikuma
6 linna (ill)
6 lähen
6 lükkas
6 lüüa
6 maailmas
6 maal (ad)
6 meenub
6 meenus
1050. 6 mehest
6 merel
6 millegi
6 millegipärast
6 mitme
6 moodi (pp)
6 mujal
6 must (a)
6 märganud
6 naer
1060. 6 neis
6 nelja (gen)
6 nendega
6 niisugused
6 niisuguseid
6 nina
6 nina (gen)
6 nähes
6 olgugi (k^{x})
6 peas
1070. 6 peast
6 pidama
6 pihku (ill)
6 piinlik
6 pikad
6 pool (n)
6 püüdsin
6 rahulikult
6 saad (v)
6 sattus
1080. 6 silmitses
6 suhtes (pp)
6 surm
6 suu (gen)
6 suust
6 taevast (part)
6 talu (gen)
6 taskus
6 tegin
6 tehtud
1090. 6 teineteisele
6 teistele (p)
6 tekkinud
6 temale
6 temalt
6 tohi (neg)
6 tuleks
6 tundub
6 tunnen
6 tähendas
1100. 6 tüdrukud

6 tühi
6 vaba
6 vaikis
6 valged
6 vanad
6 vanamehe
6 vanasti
6 vasakule (d_1)
6 vee
1110. 6 viimased
6 võeti
6 või (neg)
6 värava
6 väsinud (atr)
6 õigust
6 äärde (pp)
6 ühes
6 üht
5 aastate
1120. 5 ajaks (s)
5 ajaviiteks
5 armastus
5 astusin
5 buss
5 eemale (d_3)
5 eemalt
5 elanud
5 elust
5 emale
1130. 5 endaga
5 endine
5 enesele
5 ettevaatlikult
5 hakkama
5 hoides
5 hoidsid (pl)
5 hulgas (pp)
5 huvi (part
5 häda
1140. 5 inglise (a)
5 inimesel
5 istuma
5 juhul
5 jutu
5 jõuab
5 jõudnud (neg)
5 jäin
5 jätab
5 kakskümmend
1150. 5 kangesti
5 kartis
5 kaupa (pp)
5 keeras
5 kenasti
5 kirjutas
5 kodu (gen)
5 kokku (d_1)
5 kolmas
5 kolme (gen)
1160. 5 kombel (pp)
5 korral
5 korralik
5 kuu
5 kuulas
5 kõrged
5 käest (pp)
5 käsi
5 külas
5 küljes (pp)
1170. 5 külma (a. part)
5 laev
5 laeva (gen)
5 lapse
5 lapsepõlves
5 lasta
5 lauda (part)
5 lauset
5 leidnud
5 lendas
1180. 5 lendu (ill^x)

5 lihtsalt (d_1)
5 liikus
5 linnud
5 lõbus (a)
5 lähenes
5 lähe (neg)
5 läksime
5 meenutas
5 mehega
1190. 5 millele
5 milline
5 mitmel
5 muutusid (pl)
5 mõista
5 mõistis
5 mõistnud (neg)
5 mõnus (a)
5 mõtlema
5 mõtlen
1200. 5 märkab
5 märkamatult
5 naeratades
5 naeratas
5 nalja (part)
5 niigi (d_2)
5 niimoodi
5 niisiis
5 niivõrd
5 nimega
1210. 5 nõuab
5 näidata
5 näinud (neg)
5 näitab
5 näod
5 näost
5 nüüd (d_2)
5 nüüdki
5 olemas
5 omaette
1220. 5 otsekohe
5 paremal (d_1)
5 peaksid (pl)
5 pehme
5 perekonna
5 pidin
5 pihta (pp)
5 pikkamisi
5 pikkamööda
5 pilguga
1230. 5 polnudki
5 poole (n, gen)
5 pudeli
5 purjus
5 puu
5 puud (pl)
5 põrmugi (d_2)
5 päevad
5 päevaks
5 päikeses
1240. 5 pöördub
5 rahulik
5 rahva
5 rahvas (nom)
5 ruttas
5 rõõmus (a)
5 saama
5 saatis
5 sadama (v)
5 sakslased
1250. 5 sama (gen)
5 sama (d_1)
5 sammud (s)
5 seekord
5 seitse
5 seni (d_1)
5 siinsamas
5 silma (gen)
5 silme (pl, gen)
5 silmis
1260. 5 sinise

5 sorti (part)
5 surub
5 sõidab
5 sõitnud
5 sõnatult
5 söögilaua
5 südant
5 sügisel
5 tahaks
1270. 5 Tartu (gen)
5 teadsin
5 teiselt
5 teisest
5 teistest (p)
5 temagi
5 toime (ill^{x})
5 toimus
5 tundi (part)
5 tundsid (pl)
1280. 5 tõmbus (v)
5 tõstab
5 tähele (s^{x})
5 tänavale
5 täpselt (d_2)
5 töötle
5 tühjaks
5 unustada
5 unustanud
5 usun
1290. 5 usu (neg)
5 uued
5 vahtima
5 vaikides
5 vajub
5 valge (gen)
5 valgust (part)
5 vana (part)
5 vanamees
5 vanker
1300. 5 vastas (v)
5 vastas (pp)
5 veider
5 vene (a)
5 vihaselt
5 viis (n)
5 võetud
5 võtaks
5 väravast
5 õe
1310. 5 õhku (part)
5 õhus
5 õigemini (d_2)
5 õla
5 õlad
5 õnne (part)
5 õnnelik
5 õnnetu
5 õuel
5 öeldud
1320. 5 öö (gen)
5 üha
5 üheks
5 üksteist
4 aimata
4 aitas
4 ajast
4 aknad
4 alasti
4 algab
1330. 4 alles (d_3)
4 andma
4 armastas
4 asja (gen)
4 eemale (d_1)
4 ehk (k)
4 elasid (pl)
4 enamik
4 endamisi
4 endiselt
1340. 4 Enn

4 ennegi (d_1)
4 Ester
4 head (part)
4 heast
4 hele
4 hetke (gen)
4 hilja
4 hobuste
4 hoiab
1350. 4 hommikuti
4 hoopiski (d_2)
4 hukka (d_1)
4 huviga
4 huvitav
4 häid
4 härra
4 hüüdis
4 igas
4 imelikult
1360. 4 isaga
4 istusin
4 jalaga
4 jooksid (pl)
4 juhuslikult
4 jõudsid (pl)
4 jõuga
4 jälgib
4 jälgis (v)
4 järve (gen)
1370. 4 jää (neg)
4 kadusid (pl)
4 kaela (gen)
4 kahekesi
4 kahekümne
4 kannab
4 kartuleid
4 kasvanud
4 katki (d_1)
4 keel (nom)
1380. 4 kellest
4 kestis
4 kiita
4 kinnitas
4 kirik
4 kitsas (nom)
4 kohe (d_2)
4 kutsus
4 kuulama
4 kuuldes
1390. 4 kõigest (p)
4 kõrvalt (d_1)
4 käima
4 käisin
4 küljest (pp)
4 kümmekond
4 küsida
4 küüni (gen)
4 langeb
4 langenud
1400. 4 lapsi
4 laskus
4 lauda (gen)
4 laudu
4 leiba (part)
4 leidsin
4 leiva
4 linn
4 linnast (el)
4 loeb
1410. 4 looduse
4 lootis
4 lubanud
4 lubas
4 lume
4 lõppenud
4 lõppu (part)
4 lõunat
4 läheme
4 läinud (neg)
1420. 4 lühike

4 lükanud
4 maailm
4 madal
4 mahti (part^{x})
4 maja
4 majade
4 meeldis
4 meeldiv
4 meelest
1430. 4 merd
4 Mihkel
4 mingid (p)
4 minuga
4 muide
4 muret
4 muu (p)
4 muu (p, gen)
4 mõistsin
4 mõnd
1440. 4 mängib
4 mängis
4 märgata
4 märjaks
4 naha
4 nahk
4 naisele
4 naisi
4 nendele
4 nimesid
1450. 4 noormees
4 nõus (s/d_1)
4 nädala
4 näe (imp)
4 näe (neg)
4 näib
4 näole
4 oluline
4 ootama
4 ootasid (pl)
1460. 4 osta

4 paat (nom)
4 pakkus
4 paljajalu
4 pandud
4 pea (d_1)
4 pead (pl)
4 pelgas (v)
4 pihta (ill^{x})
4 pildi
1470. 4 pisike
4 pliidi
4 poega (part)
4 poiste
4 proovida
4 proua
4 puhas
4 puhtaks
4 põhjust
4 põlenud (atr)
1480. 4 päevade
4 päevast (el)
4 päriselt
4 pöördus
4 pühkis
4 püüdnud
4 raha
4 rahul (s/d_1)
4 riided
4 rohi
1490. 4 rõõm
4 rääkinud
4 rääkinud (neg)
4 saabus
4 saadik (pp)
4 sattunud
4 sees (pp)
4 seisin
4 seistes
4 Seiu
1500. 4 selg

4 seotud
4 silm (silma)
4 sugulased
4 suitsu (part)
4 surma (part)
4 surnud (atr)
4 surnuks
4 suuremad
4 suuremaks
1510. 4 suuri
4 suus
4 sõitis
4 sõna (gen)
4 sõnadega
4 südamesse
4 sülle
4 sündis
4 taevast (el)
4 tagasi (d_1)
1520. 4 tahtmatult
4 tahtmine
4 teisel
4 temas
4 teravalt
4 tervise
4 tihe
4 tihedalt
4 tollal
4 toolile
1530. 4 trepist
4 tuba (part)
4 tuld
4 tulnud (atr)
4 tund
4 tunni
4 tuttav (s)
4 tõmmata
4 tõtt
4 tõusta
1540. 4 tähtsam
4 tüdruku
4 uks
4 uni
4 uskunud (neg)
4 vahe (s)
4 vahib
4 vaikseks
4 vajab
4 vanaema
1550. 4 vasaku
4 vees
4 veest
4 vennad
4 vette
4 viimastel
4 viinud
4 viisakalt
4 viiskümmend
4 Villem
1560. 4 visata
4 viskas
4 voolas
4 võetud (atr)
4 võimas (a)
4 võinud (neg)
4 võtnud (neg)
4 võtsid (pl)
4 võõra (a)
4 väikest
1570. 4 väljaspool (pp)
4 õhk
4 õhtu (gen)
4 õlut
4 čues
4 äkitselt
4 öeldakse
4 ühes (p)
4 ühtegi (p, part)
4 üksikud (a)
1580. 4 üldiselt

4 üldsegi
4 ümmarguse
4 üpris
3 aastad
3 aastatel
3 abielus
3 abi (part)
3 aegade
3 ajades
1590. 3 ajalugu
3 ajanud
3 ajavad
3 alates (pp)
3 alguses (s/d_1)
3 ammugi (d_1)
3 andnud
3 antakse
3 armas
3 armastab
1600. 3 arvamus
3 arvanud
3 asetas
3 asjadest
3 asjus
3 astudes
3 astun
3 astusid (pl)
3 astusime
3 avatud (atr)
1610. 3 ehitatud
3 eit
3 elavad
3 ellu (ill^x)
3 elumaja (gen)
3 emaga
3 ema (part)
3 endal
3 ennist
3 erilist
1620. 3 esimeses
3 ette (d_1)
3 habemega
3 haigusest
3 hakkan
3 halba (s, part)
3 halli (a, part)
3 hammaste
3 head (s, part)
3 heaks (a^x)
1630. 3 heina (part)
3 Heino
3 heita
3 heledad
3 hinges
3 hirm
3 hobuseid
3 hobust
3 hoidis (v)
3 hoolega (d_1)
1640. 3 hulga (s)
3 huuled
3 huuli
3 huultega
3 huvitas
3 häbeneda
3 hämaruses
3 hävituspataljon
3 igaühel
3 iial
1650. 3 ilma (part)
3 imestasin
3 inimesena
3 isale
3 isand
3 iseenda
3 iseenesest (d_1)
3 iseennast
3 istusime
3 jaa
1660. 3 Jaani (gen)

3 jahe
3 jalga (part)
3 juhus
3 julgenud (neg)
3 jutud
3 jutustas
3 juukseid
3 jõgi
3 jõu
1670. 3 jõuame
3 jõuavad
3 jõuda
3 jõudes
3 järel (d_1)
3 järgmine
3 jätkub
3 jääma
3 jäänud (neg)
3 kaabu (gen)
1680. 3 kaduda
3 kadunud (atr)
3 kaela (ill)
3 kaela (ill^x)
3 kaetud (atr)
3 kahjuks (d_2)
3 kahtlust
3 kaldale
3 kandi (gen)
3 kannatlikult
1690. 3 kapten
3 kari (kaljurahn)
3 kartulid
3 kassid
3 kasutada
3 kasvu (part)
3 kattis
3 katuse
3 kauge
3 keha (gen)
1700. 3 keha (part)
3 kel
3 kergem
3 kerkis
3 keskel (pp)
3 kevad
3 kihutas
3 kingad
3 kiri
3 kirikus
1710. 3 kirikusse
3 kirja (gen)
3 kiskuda
3 kivi
3 kivid
3 klaasi (gen)
3 klassi (gen)
3 kleidi
3 kodu
3 koer
1720. 3 kogemata (d_1)
3 kohale (pp)
3 kohale (s/d_1)
3 kohkus
3 kollane
3 kolmanda
3 kordagi (part)
3 korraga
3 korras (s/d_1)
3 korteris
1730. 3 kostab
3 kostma
3 kuiva (part)
3 kuivanud (atr)
3 kujutas
3 kulunud (atr)
3 kumbki
3 kummaline
3 kurb
3 kuskilt
1740. 3 kutsuda

3 kuulata
3 kuulnud (neg)
3 kuulub
3 kuus (n)
3 kuuskede
3 kuuskümmend
3 kõikide
3 kõlab
3 kõndida
1750. 3 kõneles
3 kõnnib
3 kõrge
3 kõrval (d_1)
3 kõva
3 käes (s/d_1)
3 käisime
3 käivad
3 kükitas
3 külaline
1760. 3 külge (d_3)
3 külm (a)
3 küsima
3 küsisin
3 laevad
3 lagendik
3 lahtise
3 laine
3 langes
3 lasksid (pl)
1770. 3 lasti (v)
3 las
3 lauas
3 lauda (ill)
3 laul
3 laulis
3 laulsid (pl)
3 laulu (part)
3 lehma (gen)
3 lehmad
1780. 3 leidsid (pl)
3 lendasid (pl)
3 lendavad
3 leti
3 liialt
3 liigub
3 liiguvad
3 liikumist
3 lilled
3 linna (part)
1790. 3 linnas (in)
3 loomulikult (d_2)
3 lootus
3 lootust
3 lugema
3 lugenud
3 lugu (part)
3 lukus (d_1)
3 lumes
3 lõikac
1800. 3 lõpeb
3 lõpuks (d_2)
3 lõua
3 lõug
3 lõuna (gen)
3 lähedal
3 lähedalt
3 läheks
3 löönud
3 lühikese
1810. 3 maailma (part)
3 maailmast
3 maas
3 maast
3 madala
3 magama
3 magas
3 magus
3 maha (d_1)
3 maitseb
1820. 3 majja

	3	maksa (neg)		3	naerda
	3	meel		3	naerma
	3	meeldib		3	nagunii
	3	meeled		3	naisega
	3	meeles		3	naiste
	3	meenutab		3	naistest
	3	meestel		3	niipalju
	3	meetri		3	niipea (k^x)
	3	mehel		3	niisama (d_2)
1830.	3	mets	1870.	3	nime (gen)
	3	metsa (part)		3	nime (part)
	3	millestki		3	nimetas
	3	minagi		3	nimetasin
	3	mingil		3	ninaga
	3	Minna		3	nojah
	3	minul		3	noogutas
	3	minule		3	noore
	3	minus		3	number
	3	minuti		3	nurga
1840.	3	minutiga	1880.	3	nutma
	3	mis (k)		3	nõukogude (a)
	3	moel		3	nõuti
	3	muigas		3	nädal
	3	mulje		3	nädalat
	3	mure (s)		3	nähtav
	3	muudkui		3	nähtud
	3	muust		3	ollakse
	3	muutuda		3	olles
	3	mõelnud		3	olukorras
1850.	3	mõnes	1890.	3	omamoodi
	3	mõtetega		3	onu
	3	mõtte		3	oskavad
	3	mäkke		3	otsast (pp)
	3	mäleta (neg)		3	otsib
	3	märkasin		3	otsida
	3	müts		3	otsustanud
	3	müüri (gen)		3	paarkümmend
	3	naerab		3	paiku (pp)
	3	naeratab		3	paista (neg)
1860.	3	naeratus	1900.	3	paistma

3 paistsid (pl)
3 paljaks
3 paljas (nom)
3 paljude
3 paluma
3 palusin
3 pandi (v)
3 panin
3 pannud (neg)
1910. 3 paraku
3 parata
3 parema
3 paremat
3 peal (d_1)
3 pealt (d_3)
3 peamine
3 peamiselt
3 pean
3 peatus (v)
1920. 3 peeti (v)
3 peos (pihk)
3 pere (gen)
3 peremeest
3 pesu (part)
3 piki (pp)
3 pimedus
3 poe (gen)
3 poisi
3 poisike
1930. 3 poisse
3 poistega
3 pojad
3 pojale
3 poolteist
3 praegugi
3 puhtad
3 puhul (pp)
3 puid
3 punased
1940. 3 puudutas
3 põhjas (s)
3 põld
3 põldude
3 päikest
3 pööras (v)
3 pühapäeval
3 raadio
3 raamatud
3 raamatuid
1950. 3 rada (part)
3 raputas
3 rasked
3 rippusid (pl)
3 risti (d_1)
3 rohtu (part)
3 rong
3 ruttu (d_1)
3 ruumi (ill)
3 rõõmsalt
1960. 3 räägitakse
3 saadavad
3 saaks (neg)
3 saatel (pp)
3 saatma
3 saatus
3 saatuse
3 sada
3 sadakond
3 sadas
1970. 3 sagedamini
3 saigi (v)
3 Saksa (gen)
3 salaja
3 samasugused
3 sammu (gen)
3 sarnaneb
3 sauna (gen)
3 seada
3 seest (pp)
1980. 3 seetõttu

3 seinad
3 seina (gen)
3 seina (part)
3 seinu
3 seitsme
3 sekka (pp)
3 sekka (d_1)
3 seletas
3 selgemini
1990. 3 sellesama
3 sellesse
3 sellised
3 sellist (p)
3 seotud (atr)
3 serval
3 sest (p)
3 sihuke
3 siiamaani
3 sikutas
2000. 3 silla
3 silma (ill)
3 silmades
3 sinised
3 siniseks
3 sinnapoole
3 sinu
3 sirge
3 sisemine
3 sobi (neg)
2010. 3 sooja (a, gen)
3 soojust
3 suletud (atr)
3 surma (gen)
3 suudles
3 suurel
3 suurele
3 suurema
3 suurest
3 sõbra
2020. 3 sõbralik
3 sõbralikult
3 sõda (part)
3 sõit
3 sõitsid (pl)
3 sõna (part^X)
3 sõnade
3 sõrmed
3 säärast
3 sõandanud (neg)
2030. 3 sööma (v)
3 südamest
3 sügisene
3 sündmuste
3 tabas (v)
3 taga (d_1)
3 tahaksin
3 tahan
3 tahapoole
3 tahtmist
2040. 3 tahtsid (pl)
3 taipasin
3 talv (talve)
3 targem
3 tarvitsenud (neg)
3 taskust
3 teadma
3 teadmata
3 teatud (atr)
3 tegu
2050. 3 tehes
3 tehti
3 teisel (p)
3 teise (ill)
3 teisele
3 teises
3 terav
3 tige
3 tilluke
3 toast
2060. 3 tohtinud (neg)

3 toonud
3 tugevasti
3 tulgu
3 tundma
3 tungis (v)
3 tunne (neg)
3 tunned
3 tunnistada
3 tuntud (atr)
2070. 3 tuua
3 tõmbasin
3 tõusid (pl)
3 tähelepanelikult
3 tähenda (neg)
3 tähtis
3 täitsa (d_1)
3 tänava
3 tänavaid
3 tõõga
2080. 3 tüdrukut
3 tüdrukute
3 tühine
3 tütarlaps
3 uhke
3 ujus
3 uksele
3 ulatab
3 ulatas
3 ulatus (v)
2090. 3 vaata (imp)
3 vaene
3 vaevalt (d_1)
3 vahest (d_2)
3 vahtis
3 vaielda
3 vait
3 valas
3 valge
3 valguse
2100. 3 vanade
3 vanas
3 vanasse
3 vanem (a)
3 varakult
3 vargsi
3 varju (part)
3 varju (ill)
3 vasakul (d_1)
3 vastik
2110. 3 vastupidi
3 vastust
3 vedas
3 veendunud (atr)
3 veerand
3 venna
3 vennale
3 viga (part)
3 vihma (gen)
3 vihma (part)
2120. 3 viia
3 viib
3 viidi
3 viimast
3 viimaste
3 viisi (pp)
3 vili
3 viskab
3 võid (v)
3 võimatu
2130. 3 võin
3 võisin
3 vägisi
3 väikesed
3 väikesele
3 väikesi
3 vältel (pp)
3 värske
3 värvi (part)
3 väärikust
2140. 3 õhtu

3 õhtuks
3 õhtut
3 õigupoolest
3 õlgadel
3 õlgu
3 õue (part)
3 õeldes
3 üheksa
3 ühele
2150. 3 ühesõnaga
3 ühte (part)
3 üksainus
3 üksi (d_2)
3 ülal
3 ülemäära (d_1)
3 üleval
3 ümberringi
3 ütleks
3 ütlesid (pl)
2160. 2 aega (part^x)
2 aegamööda
2 aias
2 ainiti
2 ainsa
2 ainust
2 aita (neg)
2 aitab
2 aitäh
2 ajalehe
2170. 2 ajaloo
2 ajalugu (part)
2 ajama
2 akende
2 aknal
2 algavad
2 alguse
2 alustas
2 ametis
2 andnud (neg)
2180. 2 andsid (pl)
2 annan
2 annavad
2 antud (neg)
2 armastusega
2 asjadele
2 asjalik
2 astumist
2 astunud
2 asub
2190. 2 augu
2 aus (a)
2 ausalt
2 auto
2 daam
2 ebamäärane
2 ehitada
2 ehmatusest
2 ehtne
2 elaks
2200. 2 elama
2 Elmar
2 eluks
2 endasse
2 enesel
2 erakordselt
2 esile (d_3)
2 esimees
2 esimesed
2 esimesele
2210. 2 esmakordselt
2 Estri
2 foto
2 hakkaks
2 hallide (a)
2 hambad
2 hammastega
2 harjunud (atr)
2 hiljuti
2 hingab
2220. 2 hinge (part)

2 hinna
2 hirmsa
2 hirmu (part)
2 hirmus (a)
2 hirmus (d_1)
2 hirmust
2 hobusega
2 hobusel
2 hoiatas
2230. 2 hoiti
2 hommikust
2 hukkub
2 hulga (d_1)
2 hulka (pp)
2 hunnikusse
2 huulte
2 huvides
2 häiris
2 hääle
2240. 2 hääled
2 häälega
2 hüppab
2 hüüab
2 hüüda
2 hüütakse
2 iga (p, gen)
2 igast
2 igasuguseid
2 ilma (gen)
2250. 2 ilma (d_3)
2 ilma (k^x)
2 ilmaga
2 ilmusid (pl)
2 ilusaid
2 ilusasti
2 imestama
2 imestas
2 inimesega
2 inimeseks
2260. 2 inimesele
2 inimesest
2 inimestel
2 insener
2 iseendale
2 iseenesest (d_2)
2 iseloomu (gen)
2 Jaan
2 Jaanus
2 jagama
2270. 2 jagu ($part^x$)
2 jalal
2 jalgadega
2 joonud
2 Juhani
2 juhtis
2 juhtuda
2 juhuslik
2 jumala
2 jutule (s/d_1)
2280. 2 juua
2 jõud (sg)
2 jõudnud (t)
2 jäigi
2 jällegi (d_1)
2 jälle (d_2)
2 järgmise
2 järsult
2 jätkus
2 jätma
2290. 2 jätnud
2 Jüri
2 kaabu (nom)
2 Kaarel
2 kaasas (d_3)
2 kaastunne
2 kadakad
2 kadakas (nom)
2 kaetud
2 kah
2300. 2 kaheksa

2 kahele
2 kaksteist
2 kala
2 kala (gen)
2 kala (part)
2 kaldu (d_1)
2 kanda (v)
2 kaovad
2 kardab
2310. 2 kaskede
2 kasu (part)
2 kasva (neg)
2 katsuda
2 katus
2 kaugelt (d_1)
2 kaugemale (d_1)
2 kaunistatud (atr)
2 kaunis (d_1)
2 kaupa (part)
2320. 2 kaval (a)
2 kavatseb
2 kehv
2 kehvem
2 keldris
2 kerge (gen)
2 kerged
2 kergesti
2 kergitas
2 kerkib
2330. 2 kevadine
2 kiiremini
2 kilomeetri
2 kilomeetrit
2 kinni (d_1)
2 kinno
2 kirjeldada
2 kirjutada
2 kirstus
2 kiskus
2340. 2 klaasi (part)
2 kleit
2 kobas
2 kodunt
2 koera (gen)
2 koera (part)
2 kogunenud
2 kohad (koht)
2 kohal (s/d_1)
2 kohatu (a)
2350. 2 kohtusid (pl)
2 kohvikus
2 kohvikusse
2 kohvri
2 kollase
2 kollased
2 kollaseid
2 kolmekesi
2 kolmkümmend
2 kombeks
2360. 2 komme (nom)
2 konksu (gen)
2 kooli (gen)
2 koputab
2 kord (d_2)
2 koridoris
2 korra (gen)
2 korraldusi
2 korter
2 korteri
2370. 2 korterit
2 kostsid (pl)
2 kramplikult
2 kuiv
2 kujutles
2 kujutlesin
2 kukub
2 kummardus (v)
2 kunagi (d_2)
2 kurat
2380. 2 kurku (ill)

2 kusagilt
2 kutsutakse
2 kuud (part)
2 kuuleb
2 kuule (neg)
2 kuum
2 kõhn
2 kõiges
2 kõrgel (d_1)
2390. 2 kõrgemale (d_1)
2 kõrtsis
2 kõrva (gen)
2 kõrvus
2 kõõki (ill)
2 kühmus (d_1)
2 külalist
2 külasid
2 külast
2 külma (a, gen)
2400. 2 kümme
2 küsimusega
2 küsinud (neg)
2 Laas
2 labidas (nom)
2 laevas
2 lahe (laht)
2 lahku (d_1)
2 lahkuda
2 lahkuma
2410. 2 laht
2 laia (a, gen)
2 laiaõlgne
2 lained
2 laisk
2 lamas
2 langesid (pl)
2 laotud (atr)
2 laualt
2 laud (sg)
2420. 2 laudast
2 lehe
2 lehed
2 leidnud (neg)
2 leidus
2 leitud (atr)
2 libises
2 lihtsa
2 liigutada
2 liigutusega
2430. 2 liiv
2 liivale
2 lind
2 linnale
2 loevad
2 loob
2 loodus
2 loomi
2 loomulik
2 loota
2440. 2 lubada
2 lubanud (neg)
2 lumega
2 lõigatud (atr)
2 lõpetas
2 lähedal (pp)
2 lähedale (pp)
2 läheduses
2 läheks (neg)
2 lükata
2450. 2 maakera (gen)
2 maale (all)
2 madalate
2 madrused
2 magab
2 maja (part)
2 manas
2 mantli
2 masin
2 meeldi (neg)
2460. 2 meeleolu

2 meelsasti
2 meenutada
2 meestele
2 meiega
2 merele
2 merre
2 metsast
2 miilits
2 mil (millal)
2470. 2 minemas
2 mingeid
2 minusse
2 minut
2 miskipärast
2 missuguse
2 missugused
2 mitmed
2 mitmete
2 mulda (part)
2480. 2 murdunud (atr)
2 mured
2 musta (part)
2 mustad (a)
2 muusika (gen)
2 muusikat
2 muuta
2 muutunud (atr)
2 muutusin
2 muutuvad
2490. 2 mõelnud (neg)
2 mõnda (part)
2 mõnu (part)
2 mõtelda
2 mõtelnud
2 mälestus
2 mälestused
2 mäletanud (neg)
2 mäletas
2 mändide
2500. 2 mäng

2 mängus
2 männid
2 märksa (d_1)
2 määral
2 mõõdas (d_1)
2 mõõdub
2 mõõdunud (atr)
2 mütsi (gen)
2 naaber
2510. 2 naeru (gen)
2 naeru (part)
2 naisest
2 naljakas (nom)
2 neilt
2 neisse
2 neljakümne
2 niisugusel
2 nii-öelda
2 nimed
2520. 2 nimetada
2 nimetatakse
2 no (i)
2 noogutasin
2 noorem
2 noores
2 noorik
2 noormehed
2 noorte
2 normaalne
2530. 2 nukra
2 nukralt
2 nõrk
2 nõudis
2 nõutult
2 nädalad
2 nädalas
2 nägid (pl)
2 näisid (pl)
2 näitas
2540. 2 näol

2 nãol (pp)
2 nãos
2 nõõriga
2 oja (gen)
2 olegi (neg)
2 oleks (neg)
2 olete
2 olid (sg)
2 olnudki (neg)
2550. 2 oma (d_2)
2 omad (p)
2 omajagu
2 omakorda
2 ootasin
2 osanud
2 osavasti
2 oskas
2 ostetud
2 ostis
2560. 2 otsas (s)
2 otsemaid
2 otsene
2 otsinud
2 otsisid (pl)
2 otsusele (s^x)
2 paadi (gen)
2 paadis (in)
2 pagana (d_1)
2 paika (part)
2570. 2 paistnud (neg)
2 pakkuda
2 paksu (gen)
2 pane (neg)
2 parda (gen)
2 paremale (d_1)
2 parem (d_2)
2 parunite
2 pastor
2 pead (v)
2580. 2 peegli
2 peitis
2 perenaise
2 perenaisele
2 perenaist
2 perepoeg
2 peres
2 pidevalt
2 pidi (pp)
2 pigistab
2590. 2 piima (part)
2 pikali
2 pikemaks
2 pikkade
2 pikki
2 pildid
2 pileti
2 pilgud (s)
2 pilgutas
2 pilte
2600. 2 pilved
2 pinda (part)
2 pisikese
2 pistis (v)
2 pooleldi
2 pooliti
2 praegusest (a)
2 pressis (v)
2 pruukinud (neg)
2 pudelid
2610. 2 puhta (gen)
2 puhub
2 puhul (s)
2 puhus
2 purustatud (atr)
2 puudu (d_3)
2 puudub
2 puutunud (neg)
2 põlvede
2 põrandale
2620. 2 põse

2 päevi
2 päikesest
2 pärnade
2 pääse (neg)
2 põõrab
2 püha (a)
2 pükste
2 püsis
2 püüdes (v)
2630. 2 raagus
2 raamatu
2 raha (gen)
2 rahu (part)
2 rahule (s^x)
2 raskeid
2 raskelt
2 raskem
2 raskete
2 ratta (gen)
2640. 2 read
2 rebis
2 rind
2 ringi (gen)
2 rinnale
2 Robert
2 rohelised
2 rohelist
2 roninud
2 ronis
2650. 2 rutem
2 ruumi (part)
2 ruumis
2 rõõmsa
2 rätiku
2 rääkimata
2 rääkisin
2 saadab
2 saalis (s)
2 saan (v)
2660. 2 saare (veekogus)
2 said (sg)
2 sakslane
2 sakslaste
2 salli (neg)
2 sama (part)
2 samas (p)
2 sammuga
2 sarnased
2 sarved
2670. 2 seas (pp)
2 sedapuhku
2 seega (d_2)
2 seeliku
2 segada
2 segas
2 seintel
2 seisavad
2 seisnud
2 Seiu (gen)
2680. 2 seks (p)
2 selga (part)
2 selgelt
2 selgem
2 selgest
2 selgub
2 seljaga
2 selliseid
2 seltsimees
2 seob
2690. 2 siduda
2 sigareti
2 sigines
2 sihtis
2 siiani
2 siinkohal
2 siin-seal
2 sild
2 sillitas
2 sillale
2700. 2 silmadest

2 silmapilgu
2 silmil
2 silmist
2 sinine
2 sinust
2 sirutas
2 sisemiselt
2 sooja (a, part)
2 soov
2710. 2 soove
2 soovis (v)
2 soovitas
2 sulab
2 sulasid (v, pl)
2 sundis
2 surema
2 suureks
2 suuremat
2 suurt (d_1)
2720. 2 suutis
2 sõbrad
2 sõitma
2 sõitsime
2 sõja
2 sõnaga
2 sõnagi (part)
2 särav
2 säärane
2 säärase
2730. 2 sööb
2 sügava (a)
2 süüd (part)
2 tabanud
2 taevasse
2 taevas (in)
2 tagantjärele
2 tagapool (d_1)
2 tagurpidi
2 taheti
2740. 2 taipas
2 tajus (v)
2 takka (pp)
2 Tallinna (ill)
2 Tallinnas
2 talve (gen)
2 targad
2 tark
2 taustal
2 tavaline
2750. 2 tavaliselt
2 tea (imp)
2 teadus
2 teaduse
2 teatas
2 teeks (v)
2 teele (liikl.)
2 teenis
2 tegid (pl)
2 tehtud (atr)
2760. 2 teinekord
2 teisigi (p)
2 tekkinud (neg)
2 temasse
2 temperament
2 teraselt (d_1)
2 teravad
2 teravat
2 tingimata (d_2)
2 toad
2770. 2 toda
2 toetas
2 tohib
2 tohiks (neg)
2 toidu
2 toimub
2 tooma
2 toosama
2 toppis
2 tormas
2780. 2 trepil

2 tublisti
2 tugeva
2 tugevat
2 tugitoolis
2 tuhande
2 tuim
2 tuju (part)
2 tulekut
2 tuleme
2790. 2 tuli (s)
2 tulin
2 tunded
2 tundes (v)
2 tundest
2 tundide
2 tunnevad
2 tunniks
2 turi
2 tuttav (a)
2800. 2 tuult
2 tõeline
2 tõid (pl)
2 tõmbab
2 tõmbama
2 tõrjus
2 tõsise
2 tõsta
2 tõstsid (pl)
2 tähelepanu
2810. 2 tähendab
2 tähendab (d_2)
2 tähtsust
2 täitis
2 tänapäeval
2 tänavad (s)
2 tänavatel
2 tööst
2 töötab
2 töötas
2820. 2 tüdrukul
2 tühja (gen)
2 tühjus
2 tütar
2 tütarlapse
2 tütre
2 uhkus
2 uinuda
2 uksest
2 uljalt
2830. 2 und (part)
2 usku (part)
2 uskumatult
2 ust
2 vaadanud
2 vaade
2 vaatamas
2 vaatan
2 vaatasime
2 vaba (gen)
2840. 2 vabale
2 vabalt
2 vabariigi
2 vahtisin
2 vaikib
2 vaikivad
2 vaiksemaks
2 vaikust
2 valdas (v)
2 valges (a)
2850. 2 valis
2 valu (aisting)
2 valusa
2 vanaisa
2 vanaisa (gen)
2 vanameest
2 vanemad (a)
2 vao
2 varandust
2 varjata
2860. 2 varjatud (atr)

2 varjud
2 vasakut
2 vastus
2 veab
2 veeretada
2 veeretas
2 veidra
2 veri
2 viga
2870. 2 viibis
2 viie
2 viimane (s)
2 viina (part)
2 viis (s)
2 viisi (s, part)
2 vilksatab
2 voolab
2 võiksid (pl)
2 võiksin
2880. 2 võim
2 võimalikult
2 võimalus
2 võimu (gen)
2 võisid (pl)
2 võta (neg)
2 võõraks (a)
2 võõras (s, nom)
2 vägagi
2 vähemasti
2890. 2 väikeses
2 väikesest
2 väina (gen)
2 välimusega
2 välja (gen)
2 väravas
2 õhku (ill)
2 õhtuni
2 õigel
2 õiglane
2900. 2 õnn
2 õpetaja
2 õue (gen)
2 õue (ill)
2 õueväravas
2 äkki (d_2)
2 ämber
2 ärgu
2 äärt
2 ööd (pl)
2910. 2 ööd (part)
2 öösiti
2 üheksateistkümne
2 üheskoos
2 ühine
2 üht (p)
2 ühtemoodi
2 ühtlasi (d_1)
2 üht-teist
2 üksik (a)
2920. 2 üksikuid (a)
2 üksteisest
2 ülejäänud (atr)
2 üleliia
2 ülevalt
2 ülikoolis
2 ümbert (pp)
2 ümbruskonna

Viidatud kirjandus

A l l é n, S., Nusvensk frekvensordbok, baserad på tidningstext. 1. Graford, homografkomponenter. Stockholm, Almquist & Wiksell, 1970.

M i s t r í k, J., Frekvencia slov v slovenčine. Bratislava, 1969.

J u i l l a n d, A., C h a n g - R o d r i g u e z, E., Frequency Dictionary of Spanish Words. The Hague, Mouton, 1964.

J u i l l a n d, A., E d w a r d s, P. M. H., J u i l l a n d, I., Frequency Dictionary of Rumanian Words, The Hague, Mouton, 1965.

J u i l l a n d, A., B r o d i n, D., D a v i d o v i t c h, C., Frequency Dictionary of French Words. The Hague, Mouton, 1970.

R ä t s e p, H., Eesti keele väljendverbide olemusest. - "Keel ja Kirjandus", 1973, nr. 1, lk. 24-30.

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.

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

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

А л е к с е е в П. М. Некоторые вопросы теории и практики статистической лексикографии. - Статистика текста. Т. 1. Минск, Изд. БГУ, 1969, 12-37.

Д а р ч у к Н. П. Статистические параметры лексики в произведениях Т. Смолича и Г. Тютюнника. - Вопросы статистической стилистики. Киев, "Наукова думка", 1974, 262-275.

ЧАСТОТНЫЙ СЛОВАРЬ СЛОВОФОРМ АВТОРСКОЙ РЕЧИ ЭСТОНСКОЙ ХУДОЖЕСТВЕННОЙ ПРОЗЫ

Ю. Каазик, Ю. Тулдава, А. Виллуп, К. Ээремаа

Р е з ю м е

В статье публикуется первая часть частотного словаря авторской речи современной эстонской художественной прозы – словарь словоформ. Выборка состоит из 20 подвыборок по 5000 словоупотреблений из разных произведений, опубликованных после 1960 года. Словарь охватывает 30733 словоформы, из них 21760 встречались только по одному разу (что составляет 70,8% объема словаря и 21,8 % объема текста). В статье публикуются 3000 самых частых словоформ, расположенных по порядку убывания так называемых "модифицированных" частот, которые вычисляются по формуле:

$$U = D \cdot F$$

где U – модифицированная частота (коэффициент употребления), F – абсолютная частота словоформы в тексте, D – коэффициент стабильности (по А. Кюйану):

$$D = 1 - \frac{v}{k - 1}$$

(v – коэффициент вариации, k – число подвыборок).

В словаре различаются грамматические и лексические синонимы.

Словарь лексем (алфавитный и частотный) и более подробный анализ материалов словаря будут опубликованы в следующем выпуске сборника.

Работа по составлению и обработке частотного словаря выполнена членами Группы лингвостатистики филологического факультета и сотрудниками Вычислительного центра Тартуского государственного университета.

A FREQUENCY DICTIONARY OF MODERN ESTONIAN PROSE FICTION

Ü. Kaasik, J. Tuldava, A. Villup, K. Käremaa

Summary

The article presents the first part of a frequency dictionary of modern Estonian prose fiction (non-conversational material) - a dictionary of forms. The sample consists of 100,000 running words (20 subsamples of 5,000 words each) from various works published after 1960. The dictionary includes 30,733 words (forms), of which 21,760 occurred once only (which is 70.8 % of the volume of the dictionary and 21.8 % of that of the text). In this article the 3,000 most frequent words (forms) are published in the order of so-called "modified" frequency calculated according to the formula:

$$U = D \cdot F$$

where U - modified frequency (coefficient of usage), F - absolute frequency of the word in a text, D - coefficient of distribution (according to A. Juilland):

$$D = 1 - \frac{v}{k - 1}$$

(v - coefficient of variation, k - number of subsamples). A distinction is made between grammatical and lexical homonyms.

A dictionary of lexemes (alphabetical and frequency list) and a more detailed analysis of the materials on the vocabulary will be published in the next issue of the present series.

The frequency dictionary has been compiled and materials elaborated by the members of the Group of Linguostatistics and the Computing Center of the Tartu State University.

SISUKORD - CONTENTS

J. Tuldava, Statistilised meetodid ja keeleteadus. - Statistical Methods and Linguistics (in Estonian) 5- 59
Summary in English 60

J. Tuldava, A. Villup, Sõnaliikide sagedusest ilukirjandusproosa autorikõnes. - Statistical Analysis of the Parts of Speech in Estonian Fiction (in Estonian). 61-102
Summary in English 105-106

Ü. Kaasik, J. Tuldava, A. Villup, K. Ääremaa, Eesti keele ilukirjandusproosa autorikõne sagedussõnastik. - Frequency Dictionary of Estonian Fiction (Word-Forms in Non-Conversational Materjal) 107-151
Summary in English 153

СОДЕРЖАНИЕ

Ю. Тулдава. Статистические методы и языкознание (на эст. яз.) 5- 59
Резюме (на рус. яз.) 60

Ю. Тулдава, А. Виллуп. О частотности частей речи в авторской речи художественной прозы (на эст. яз.) 6I-I02
Резюме (на рус. яз.) I03-I04

Ю. Каазик, Ю. Тулдава, А. Виллуп, К. Ээрема. Частотный словарь словоформ авторской речи эстонской художественной прозы (на эст. яз.) I07-I5I
Резюме (на рус. яз.) I52

ТРУДЫ ПО ЛИНГВОСТАТИСТИКЕ I. Ученые записки Тартуского государственного университета. Выпуск 377. На эстонском и русском языках. Резюме на русском и английском языках. Тартуский государственный университет. ЭССР, г. Тарту, ул. Юликооли,I8.

Vastutav toimetaja J. Soontak. Paljundamisele antud 23.XII.75.Trükipaber nr. 1. 30x45 1/4.Trükipoognaid 9,75. Arvestuspoognaid 9,15. Trükiarv 500.MB 08345. TRÜ trükikoda, ENSV, Tartu, Pälsoni t.14. Tell. nr. 1479. Hind 91 kop.