SAGEDUSSÕNASTIK LEKSIKOSTATISTILISE UURIMISE OBJEKTINA

J. Tuldava

Igasuguse leksikostatistilise uurimise aluseks on vastava keele või allkeele põhjal koostatud sagedussõnastik. Alljärgnevas artiklis vaatleme sagedussõnastike koostamise põhimõtteid ja, toetudes eesti ilukirjandusproosa autorikõne sõnavormide ja lekseemide sagedussõnastikele, esitame mõningaid praktilisi ja teoreetilisi järeldusi eesti keele sõnavara statistilise struktuuri kohta.+ Eraldi võtame vaatluse alla sellised olulised sõnavarastatistilised seaduspärasused nagu Zipfi seadus sõnasageduse ja astaku vahekorrast ning sõnasageduste jaotumuse (leksikaalse spektri) teooria, illustreerides teoreetilist materjali näidetega eesti keelest.

Sagedussõnastike koostamise põhimõtted. Teatavasti on käesolevaks ajaks koostatud sagedussõnastikud peaaegu kõigi kultuurkeelte kohta. Olemasolevatel andmetel läheneb sagedussõnastike arv kogu maailmas 400-le. Ainuüksi Nõukogude Liidus on loodud umbes 100 sagedussõnastikku enam kui 20 keele kohta (sealhulgas ka võõrkeelte ja mitmesuguste tehniliste allkeelte kohta). Sagedussõnastiku koostamine on töömahukas protsess, milles tavaliselt osalevad suuremad kollektiivid (nii näit. tegeleb läti keele sagedussõnastiku koostamisega 10-liikmeline rühm - lingvistid, matemaatikud, abitöötajad -, kusjuures kasutatakse ulatuslikult elektronarvuti abi). See töö tasub aga ennast ära, sest sagedussõnastikud on muutunud hädavajalikuks vahendiks nii lingvistilises kui ka pedagoogilises ja psühholoogilises uurimistöös. Sagedussõnastikud on aluseks miinimumsõnastike, mitmesuguste tõlke- ja terminoloogiasõnaraamatute koostamisel, samuti side, stenograafia, masinatõlke jm. probleemide lahendamisel. Tänapäeval, mil sagedussõnastikke koostatakse peamiselt elektronarvutite kaasabil, võib uurimistöö osutuda eriti viljakaks sel juhul, kui sagedussõnastik ja selle aluseks olev tekst säilitatakse arvutis, mis võimaldab korduvat pöördumist nii sõnastiku kui teksti poole uute esilekerkivate probleemide lahendamiseks.

Sagedussõnastikud jagunevad mitmetesse eri liikidesse.

S u u r u s e järgi eristatakse väikseid sagedussõnastikke (vastava tekstimahuga kuni 500.000 sõnet), keskmise suurusega (1 - 10 miljonit sõnet) ja suuri sagedussõnastikke (üle 10 miljoni sõne). Suuri sagedussõnastikke on käesolevaks ajaks koostatud ainult mõned üksikud, näit. F. Kaedingi (1898) saksa keele sagedussõnastik tekstimahuga 11 miljonit sõnet, E. Thorndike'i ja J. Lorge'i inglise keele sagedussõnastik tekstimahuga 18 miljonit sõnet (Thorndike,

+ Osa vajalikke kommentaare eesti tänapäeva ilukirjandusproosa autorikõne sagedussõnastiku kohta leiab lugeja käesoleva kogumiku sissejuhatavast artiklist.

Lorge, 1944). Ka keskmise suurusega sõnastikke pole kuigi palju, näitena võib tuua 7 miljonil sõnel põhineva poola keele vanema perioodi sagedussõnastiku (Słownik, 1956), tšehhi keele sagedussõnastiku 1,2 miljoni sõnega (Bečka, Jelínek, Těšitelová, 1961) ja rida 1.000.000-sõnelisi sagedussõnastikke (Josselson, 1953; Kučera, Francis, 1967; Mistrík, 1969; Allén, 1970; Засорина, 1977). Suurem osa olemasolevaist sagedussõnastikest kuuluvad väikeste sõnaraamatute hulka, kuid ka need suudavad küllaldase usaldatavusega esile tuua keele põhisõnavara.

Sagedussõnastike koostamisel on oluline arvestada tekstide d o s e e r i m i s t allkeelte järgi. Paljud keelestatistikud on arvamusel, et tõelist väärtust omavad vaid allkeelte sagedussõnastikud, mis põhinevad enam-vähem ühtlasel (homogeensel) keelematerjalil. On koostatud suuremaid sagedussõnastikke ainult teatava allkeele kohta, näit. rootsi ajalehekeele sagedussõnastik tekstimahuga 1 miljon sõnet (Allén, 1970), ukraina ilukirjandusproosa sagedussõnastik tekstimahuga 400.000 sõnet (Перебейнос, 1969). Eriline koht on tekstide automaattöötluse - masinatõlke, infootsi- jm. probleemide lahendamiseks mõeldud kitsaste allkeelte (näit. elektroonika- või raadioalaste tekstide) sagedussõnastikel, mida on viimastel aastatel hulgaliselt koostatud Nõukogude Liidus (näit. Алексеев, 1965; Калинина, 1968; Зореф, 1971). Selliste sõnastike põhjal on loodud ka pedagoogilise suunitlusega miinimumsõnastikke inglise, saksa jt. keelte alal (Ротарь, 1970; Алексеев, Турыгина, 1974).

Üldkeele sagedussõnastike koostamisel kehtivad kokkuleppelised nõuded, kuid võib täheldada suuri erinevusi eri keelte sagedussõnastike puhul. Nii näiteks on H. Kučera ja W. Francise (1967) ameerika inglise keele sagedussõnastikus allkeeled esindatud järgmistes proportsioonides: ilukirjandusproosat 60 %, luulet 10 %, ajalehetekste 20 % ja teadustekste 10 %. J. Mistríki (1969) slovaki sagedussõnastikus esinevad allkeeled järgmises vahekorras: ilukirjandusproosa 30 %, draamadialoog 10 %, luule 13 %, ajalehetekstid 15 %, teadustekstid 32 %. Eeskujuliku metoodika alusel koostatud prantsuse keele sagedussõnastik (Juilland jt., 1970) põhineb viiel allkeelel, kusjuures iga allkeel on esindatud 100.000-sõnelise tekstiga (ilukirjandusproosa, draama, essee, ajalehetekstid, teadustekstid). Ka hispaania keele sagedussõnastiku (García Hoz, 1963) aluseks on 100.000-sõnelised tekstid eri allkeeltest (ilukirjandus, ajalehed, ametlikud dokumendid, erakirjavahetus). Soome keele sagedussõnastik (Saukkonen, 1975) hõlmab neli allkeelt (teatmekirjandus, raadiosaated, ajalehed, ilukirjandus), kusjuures iga allkeele tekstide maht on samuti 100.000 sõnet. Hiljuti valminud uues vene kirjakeele sagedussõnastikus, mille üldmaht haarab 1 miljon sõnet (Засорина, 1977) on neli allkeelt esindatud võrdselt (ilukirjandusproosa, realistlik draama, teadusalane publitsistika ja ajalehed). Läti keele sagedussõnastik on moodustatud kolme allkeele baasil (ilukirjandus, ajalehe-, teadustekstid), kusjuures iga allkeel on esindatud 300.000-sõnelise tekstiga. Läti keele sagedussõnastikus on ilukirjandus jaotatud omakorda kolme ossa (proosa, draama, luule) tekstimahuga à 100.000 sõnet. Võiks tuua veel palju näiteid erinevatest lahendustest sagedussõnastike koostamisel, kuid

üldiselt kehtib reegel, et andmed esitatakse nii allkeelte kui ka "üldkeele" kohta eraldi.

Keelte või allkeelte sagedussõnastike kõrval on tuntud ka nn. k i r j a n i k u s õ n a s t i k u d, milles esitatakse sõnade sagedusloendid ühe kirjaniku või osa teoste alusel. Nii näiteks on koostatud kirjanikusõnastik A. Puškini teoste põhjal tekstimahuga 544.777 sõnet ja sõnastikumahuga 21.197 lekseemi (vt. Фрумкина, 1961). Ka Rootsis ilmunud eesti keele sagedussõnastik (Tauli, 1964) on tegelikult kirjanikusõnastik (koostatud A. Mälgu romaani "Tee kaevule" I põhjal).

Sagedussõnastiku koostamisel aluseks võetud keeleliste üksuse järgi eristatakse sõnavormide, lekseemide, sõnaühendite, häälikuühendite ja sõnaosade (silpide, morfeemide) jm. sõnastikke. Sünteetiliste keelte puhul võivad s õ n a v o r m i d e ja l e k s e e m i d e sagedussõnastike vahel olla küllaltki suured erinevused. Sõnavara kvantitatiivsel uurimisel on aga mõlemat tüüpi sõnastikud vajalikud. Sõnavormide sõnastik võimaldab täielikumalt haarata uuritavat tekstimaterjali, sest see peegeldab ka statistilis-morfoloogilist tekstistruktuuri. Sõnavormide sõnastiku abil võib analüüsida grammatiliste vormide esinemust ja paremini kindlaks määrata sõnavara välisparameetrid (foneetilis-grafeemilised mõõted, sõnapikkused jm.), mis on vajalikud tekstide automaattöötluse probleemide lahendamisel, sealhulgas ka elektronarvuti mälu ratsionaalsemal kasutamisel. Võib öelda, et sõnavormide sagedussõnastikku saab kasutada suurema arvu ülesannete lahendamisel kui lekseemide sõnastikku, pealegi on alati võimalik moodustada sõnavormide sõnastikust lekseemide sõnastik, samal ajal kui vastupidine ei ole võimalik. Lekseemide sõnastik on vajalik üldisemat tüüpi leksikoloogiliste ülesannete lahendamisel. Eriti kasulik on kombineeritud lekseemide-sõnavormide sõnastik, kus lekseemide juures on antud kõik tekstides esinenud grammatilised vormid (selline kompleksne sagedussõnastik oleks aga väga mahukas, eriti sünteetiliste keelte, näiteks eesti keele puhul).

S õ n a ü h e n d i t e sagedussõnastikud võivad tuua suurt kasu fraseoloogia uurimisel ja õpetamisel (näit. saksa keele idioomide sagedussõnastik, vt. Hauch, 1931; inglise keele sõnaühendite sagedussõnastik, vt. Частотный словарь, 1972). Sõnaühendite sagedussõnastikke, eriti nn. triaadide sõnastikke (vt. näit. Данейко и др., 1968) kasutatakse ka tekstide automaattöötluse huvides. Morfeemide sagedussõnastikke (näit. Пиквер, 1976) võib rakendada nii tavalises lingvistilises analüüsis kui ka tekstide automaatsegmenteerimisel.

Keeleõppimise seisukohalt on eriti vajalikud s e m a n t i l i s e d sagedussõnastikud, kus eristatakse sagedusi sõna eri tähenduste alusel. Tuntuim on M. Westi inglise keele semantiline sagedussõnastik (West, 1953). Seda tüüpi sõnastiku eriliik on nn. asendussõnastik, milles võetakse arvesse nii sõnade eri tähendusi kui ka sünonüümiat (Ратцева, Строганов, 1974). Nn. t r a n s s e m a n t i l i n e sagedussõnastik ehk tõlkesagedussõnastik sisaldab võrdlevaid andmeid kahe või enama keele kohta, näit. H. Eatoni inglise-prantsuse-saksa-hispaania sagedussõnastik (Eaton, 1961). Transsemantilisi automaatsõnastikke kasutatakse käesoleval ajal masinatõlke-, infoot-

si- jm. probleemide lahendamisel (vt. Вертель и др., 1971).

Igasugune sagedussõnastik võib esineda kahel põhilisel kujul: tähestikulise loendina või sagedusloendina. T ä h e s t i k u l i s e s l o e n d i s (loetelus) esitatakse sõnad või sõnavormid alfabeetilises järjestuses koos sagedusandmetega. S a g e d u s l o e n d i s on sõnad või sõnavormid järjestatud sageduste kahanevas reas. Mõlemal juhul võivad sõnade (sõnavormide) koondsageduse kõrval esineda ka sagedused osavalimite kaupa (viimasel juhul nimetatakse loendit "jaotussõnastikuks"). Tähestikulise ja sagedusloendi kõrval esinevad ka mitmesugused lisaloendid, näit. pärisnimede, hübriidsõnade, logogrammide (sümbolite, märkide) jm. loendid koos sagedusandmetega. Kui sõnastikus on ära toodud kõik laused või lauseosad, kus vastav sõna esineb, ning ära märgitud vastav teos (ja leheküIg), siis nimetatakse sellist loetelu k o n k o r d a n t s s õ n a s t i k u k s. Konkordantssõnastikud on eriti vajalikud filoloogilises ja kirjandusteaduslikus ning stilistilises uurimistöös. Viimasel ajal on selliseid sõnastikke hulgaliselt koostatud elektronarvutite kaasabil (vt. näit. Виберштейн, 1971; Engwall, 1974; Papp, 1975).

Mitmesuguse morfoloogia-alase uurimistöö hõlbustamiseks on hakatud kasutama p ö ö r d s õ n a s t i k k e, mis võivad olla varustatud sagedusandmetega. Pöördsõnastikus asetatakse sõnad (või sõnavormid) alfabeetilisse järjekorda sõna lõpust loetavate tähtede järgi. Esimesed pöördsõnastikud koostati juba 1904.a. ladina ja vana-iraani keele kohta (vt. Štindlová, 1960). Suuremal arvul hakati pöördsõnastikke moodustama viimasel aastakümnel, mil avanes võimalus kasutada elektronarvuti abi. Pöördsõnastikud on olemas muuseas ka läti, soome, ungari ja eesti keele kohta. Esimene eesti keele pöördsõnastik koostati SFV-s R. Hinderlingi poolt "Õigekeelsuse sõnaraamatu" põhjal sõnade põhivormide alusel ilma sagedusandmeteta (vt. Hinderling, 1975). TRÜ arvutuskeskuse ja filoloogiateaduskonna rakenduslingvistika rühma ühistööna on valminud sagedusandmetega varustatud eesti keele pöördsõnastik (s.o. pöördsagedussõnastik) ilukirjandusproosa autorikõne tekstide põhjal sõnavormide tasandil (avaldatakse edaspidi).

Sagedussõnastikke võib avaldada täielikult või osaliselt. Et madala sagedusega sõnade statistiline usaldatavus on tavaliselt väike ja nende kvantitatiivset üleolekut paljudest teistest, seega sagedussõnastikku mitte sattunud sõnadest ei saa tõestada, siis jäetakse niisugused sõnad sageli loetelust välja. Sel teel saavutatakse ka sõnastiku mahu kokkuhoid, kusjuures põhisõnavara on siiski haaratud. Mitmesuguste leksikoloogiliste ja stilistiliste ülesannete lahendamiseks on aga kasulik omada andmeid kõigi vaadeldavate tekstide koondsõnavara kohta. Nimetasime juba eelpool, et elektronarvutite kaasabil koostatavate sagedussõnastike materjalid (lähtetekstid, osavalimite sõnastikud jm.) säilitatakse tavaliselt arvutis, mis võimaldab jatkuval uurimisel korduvalt pöörduda algandmete poole. Eesti ilukirjandusproosa autorikõne tekstid ja sagedussõnastiku mitmesugused variandid on hoiul TRÜ arvutuskeskuses ja rakenduslingvistika rühmas.

Eesti keele sagedussõnastik. Esimene suurema ulatusega eesti keele sagedussõnastik koostati TRÜ arvutuskes-

kuse ja rakenduslingvistika rühma ühistööna 1975.a. Sellest on avaldatud esimene osa - tänapäeva ilukirjandusproosa autorikõne sõnavormide sagedussõnastik (3000 sagedama sõnavormi ulatuses) - kogumiku "Keelestatistika" esimeses väljaandes (Kaasik, Tuldava, Villup, Ääremaa, 1976). Kogumiku käesolevas (teises) väljaandes avaldatakse ilukirjanduproosa autorikõne lekseemide sagedussõnastik täielikult koos mitme lisaloendiga. Esialgselt on eesti keele sagedussõnastik kavandatud nelja allkeele kohta (ilukirjandusproosa autorikõne, tegelaskõne ja draamadialoog, ajalehekeel, teaduskeel) ning kokkuvõtlikult üldkeele kohta. Igast allkeelest võetakse 100.000-sõneline valim, kusjuures koguvalim (koondvalim) koosneb parema representatiivsuse tagamiseks paljudest osavalimitest.

Esimese allkeelena on käesolevaks ajaks statistiliselt töödeldud eesti tänapäeva ilukirjandusproosa a u t o r i k õ n e (autoritekst). Ilukirjandusproosa autorikõne valimine esimeseks uurimisobjektiks pole juhuslik. Paljude uurijate arvates on ilukirjanduslikul autoritekstil eriline tähendus teiste allkeelte tekstide seas. Autorikõnes väljendab kirjutaja enda mõtteid kõige otsesemalt ning tekstide temaatiline ja stilistiline diapasoon on kõige ulatuslikum (vrd. Bamak, 1974, 323). Ilukirjanduslik autoritekst saavutab seetõttu mõneski mõttes keskse koha antud keele tekstide hulgas. Uurimused on näidanud, et mitmete parameetrite poolest läheneb ilukirjandusproosa autorikõne kõige enam üldkeele keskmistele näitajatele. See ei tulene mitte ainult sellest, et "üldkeele" konstrueerimisel on senistes uurimustes ilukirjandustekstil määrav osa olnud (tavaliselt 25-50 %). Oluline on see, et ilukirjandusproosa autorikõne suhtes võib konstateerida suurimat interkorrelatsiooni teiste allkeeltega mitmetel eri tasanditel (näiteks ka eesti keele grafeemostatistiliste mõõdete võrdlemisel, vt. Tuldava, 1976). Võib teha järelduse, et ilukirjandusproosa autoriteksti alusel saab kindlaks määrata mitmed üldkeelele iseloomulikud leksikostatistilised omadused (teksti genereerimise üldised kvantitatiivsed seaduspärasused, üldkasutatav põhisõnavara jm.). Ilukirjandusproosa autorikõne võib antud etapil kõige paremini esindada eesti üldkeelt ja autorikõne uurimisel saadud kogemusi, järeldusi ja meetodeid võib edaspidi üle kanda kõigi teiste allkeelte lingvostatistilisse uurimisse.

Tänapäeva eesti ilukirjandusproosa autorikõne sagedussõnastik põhineb tekstidel, mis on võetud pärast 1960.a. ilmunud teostest. Koondvalim koosneb 20 osavalimist à 5000 sõnet kaheksümnest eri teosest (lähemalt vt. Kaasik, Tuldava, Villup, Ääremaa, 1976, 108). Iga osavalim jaguneb omakorda viieks juhuslikult valitud tekstilõiguks à 1000 sõnet teose erinevatest osadest. Algtekstid analüüsiti läbi grammatiliselt ja homonüümid indekseeriti. Seejuures kasutati indekseerimissüsteemi, mille kohaselt sõnavormidele lisati (leksikaalsete ja grammatiliste homonüümide korral) kaks arvu järgmise tabeli alusel:

Arv	Esimese arvuna	Teise arvuna
0	pärisnimi	(vaba)
1	1. tähendus	nimetav käände
2	2. tähendus	omastav käände
3	3. tähendus	osastav käände
4	nimisõna	muu käände

5	omadussõna	ainsus
6	asesõna	mitmus
7	tegusõna	ühend- või väljendverbi määr- või käändsõnaline komponent
8	määrsõna	eitus; rõhumäärsõna
9	sidesõna	käskiv kõneviis

Indeksi esimene arv näitab sõna tähendust või sõnaliiki, teine arv aga käändeid või muid grammatilisi vorme. Näit. sõnavorm tööd võis esineda kujul tööd45 või tööd46: esimesel juhul on tegemist ainsusega, teisel juhul mitmusega. Käände märkimine pole siin oluline, sest ainsuse korral saab antud sõnavormi puhul tegemist olla ainult osastava käändega, mitmuse korral aga nimetavaga. Sõnavorm töö märgiti aga töö41 või töö42, olenevalt sellest, kas see esines nimetavas või omastavas käändes. Indekseerimisel võeti arvesse mõned lisatingimused, nii näiteks jäeti märkimata kaassõna, kuna aga homonüümsed määrsõnad märgiti (üle, alla, sisse jt.). Eri tähendusega homonüümide järjestamisel arvestati "Õigekeelsuse sõnaraamatus" (1960) toodud järjekorda, näit. tee (jook) ja tee (liikl.) indekseeriti vastavalt tee11 (või tee12 - omastava käände puhul) ja tee21 (või tee22). Verbina võis tee esineda kujul tee78 ('ei tee') või tee79 ('tee!'). Eristati ka keskõnu täiendi või öeldistäite funktsioonis, s.o. nominaalses funktsioonis ja tarindina (tähistused vastavalt 50 ja 55). Kokkuvõttes osutus meie indekseerimissüsteem (koos mõningate kokkuleppeliste lisatingimustega) täiesti küllaldaseks ära märkimaks homonüümiajuhud eestikeelses tekstis masintöötluse otstarbeks.

Pärast tekstide eeltöötlust (nn. "prepareerimist") perforeeriti materjal viierajalisele lindile. Edasi toimus töötlemine juba automaatselt - elektronarvuti abil. Perforeeritud materjali põhjal koostas arvuti sõnavormide sagedussõnastiku mitmesugused variandid (vt. Kaasik, Tuldava, Villup, Ääremaa, 1976, 109). Hilisem lemmatiseerimine (sõnavormide koondamine põhivormi alla) toimus käsitsi. Selle tulemusel saadi lekseemide sagedussõnastik.

Eesti tänapäeva ilukirjandusproosa autorikõne sagedussõnastiku põhiparameetrid on avaldatud käesoleva kogumiku avaartiklis, seepärast me siinkohal nendel ei peatu. Kokkuvõtlikud statistilised andmed sagedussõnastiku struktuuri kohta nii sõnavormide kui lekseemide tasandil esitame artikli lõpus tabelites 1 ja 2. Keelte tüpoloogilise uurimise seisukohast pakub huvi tekstide katmuse võrdlemine s õ n a v o r m i d e tasandil. Uurimused on avastanud, et indoeuroopa keeltes katavad poole tekstist keskmiselt 80-200 sagedamat sõnavormi, kuna aga näiteks türgi-tatari keeltes läheb selleks vaja 700-800 sõnavormi (Вектаев, 1972, 9).+ Meie sagedussõnastiku andmetel katavad poole tekstist ligikaudu 900 sagedamat sõnavormi (vt. tabel 1). Eesti keel seisab selles mõttes türgi-tatari keeltele tunduvalt lähemal kui indo-euroopa keeltele, mis on seletatav keele morfoloogilise ehituse iseärasustega. Kvalitatiivseid omadusi

+ Tingimuseks on, et vaadeldava teksti maht oleks küllalt suur. Rootsi keelestatistiku S. Alléni (1970a) uurimuste põhjal saavutavad tekstikatmuse näitajad stabiilsuse alates N ≥ 20.000 (s.o. teksti pikkus vähemalt 20.000 sõnet).

registreerib tundlikult kvantitatiivne näitaja ja nii on põhimõtteliselt võimalik mingi tundmatu keele teksti puhul automaatselt kindlaks määrata keeletüüp. Neid küsimusi on põhjalikult uurinud K. Bektajev oma hiljuti kaitstud doktoritöös (Бектаев, 1975). Esitame K. Bektajevi skeemi keeletüüpide määramiseks 100 sagedama sõnavormi tekstikatmuse alusel:

aglutinatiivsed keeled	20–28 %,
sünteetilis-flektiivsed keeled	24–42 %,
analüütilis-flektiivsed keeled	43–54 %,
analüütilis-flektiivsed keeled amorfsuse elementidega	48–60 %.

Eesti keel (vastav protsent 28,9) satub sünteetilis-flektiivsete keelte rühma, olles siiski lähedane ka aglutinatiivsetele keeltele. Alljärgnevas tabelis esitame kõrvutavad andmed teksti katmuse kohta sõnavormidega (protsentides) mitmes eri keeles (Перебейнос, 1969; Ротарь, 1970; Allén, 1970; Kučera, Francis, 1967):

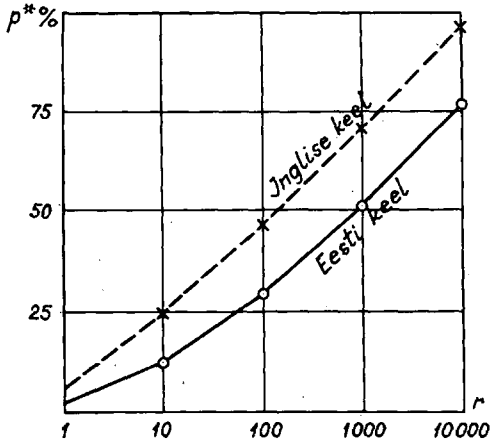
Astak:	10	50	100	1000	2000
Eesti keel	12,9	23,2	28,9	51,2	58,9
Ukraina keel	13,5	25,9	32,0	52,4	58,5
Rootsi keel	20,2	37,7	43,9	63,2	69,3
Saksa keel	18,1	-	48,7	70,5	78,9
Inglise keel	24,5	40,8	47,6	70,5	79,0
Prantsuse keel	28,2	-	57,9	79,1	85,6

Ilmnevad suured erinevused analüütiliste ja sünteetiliste keelte vahel. Kui näiteks inglise ja prantsuse keeles katavad 2000 sagedamat sõnavormi vastavalt 79 ja 86 % tekstist, siis on eestikeelse teksti puhul vastav protsent ainult 59. Esitatud keeletüpoloogilised andmed omavad praktilist väärtust keeleõpetuses (õppetekstide sõnavara doseerimisel, miinimumsõnastike koostamisel jne.) ning mitmesuguste muude rakenduslingvistiliste probleemide lahendamisel, sealhulgas ka tekstide automaattöötluse huvides. Joonisel 1 võib jälgida sõnavormide tekstikatmuse dünaamikat kõrvutavalt eesti ja inglise keeles (ühesuuruste 100.000-sõneliste valimite puhul).

Ka l e k s e e m i d e tasandil on erinevused tekstikatmuses küllaltki märgatavad eriti kõrgsagedusega sõnade osas. 10 sagedamat lekseemi katavad eesti keeles 18–20 % tekstist, inglise ja prantsuse keeles aga 26–30 % (vt. Алексеев, Турыгина, 1974; Engwall, 1974). 50 sagedama lekseemi tekstikatmus on eesti keeles 33 %, inglise ja prantsuse keeles aga 43–46 %. Erinevused tulenevad peamiselt sellest, et analüütilistes keeltes kasutatakse palju kõrgsagedusega artikleid ja eessõnu, mis sünteetilises keeles tavaliselt puuduvad. Võrrelgem näit. 10 sagedamat lekseemi eri keeltes:

eesti keeles: olema, ja, ta, see, ei, et, kui, mis, ma, tema;
inglise keeles: the, be, of, a (an), and, in, to, have, for, that;
prantsuse keeles: le, de, je, être, avoir, un, et, à, qui.

Andmed eri keelte tekstikatmuse kohta lekseemide tasandil lähenevad üksteisele, kui haarata kaasa keskmise sagedusega sõnad. Võib nõustuda R. Frumkina väitega



Joon. 1. Tekstikatmuse dünaamika eesti ja inglise keeles: r - sõnavormi astak, P^+- tekstikatmus. Poollogaritmdiagramm.

(Фрумкина, 1961), mille kohaselt suuremas osas keeltes (olenemata keeletüübist) peaks 1500 sagedama lekseemi tekstikatmus kõikuma 80 ± 10 % ümber. Meie sagedussõnastiku andmeil moodustavad 1500 sagedamat lekseemi eestikeelsest tekstist ligikaudu 77 % (vt. tabel 2).

Sagedussõnastiku usaldatavus. Sagedussõnastiku usaldatavuse all mõeldakse traditsiooniliselt sõnasageduste ja -astakute täpsust sõnastikus, s.o. nende lähedust tõenäosuslikele väärtustele. Antud juhul tekib küsimus meie sagedussõnastikus esinevate sõnade sageduste usalduspiiridest ja astakute vastavusest tegelikkusele. Eeldades, et kõrg- ja keskssagedusega sõnad jaotuvad ligikaudu normaalselt või Poissoni jaotuse järgi (statistilistest jaotustest keeleteaduses lähemalt vt. Tuldava, 1976a) ja võttes sageduste hajuvuse hindamisel aluseks küllaltki range representatiivsuse vea (esindusvea) hinnangu $\pm t \sqrt{F}$, võime kindlaks teha sõnasageduste usalduspiirid antud usaldusnivool. 95 %-lise usaldusnivoo (statistilise kindluse) korral on konstandi t väärtuseks 2. Kõige sagedama sõna (lekseemi) olema puhul (sagedus F = 4237) saame esindusvea suuruseks $2\sqrt{4237} = 130$. Sõna sageduse usalduspiirid on seega määratud väljendiga 4237 ± 130, s.o. 4107 ... 4367 (protsentides moodustab see 4,1 ... 4,3 % tekstist). Nii edasi arvutades saame sagedussõnastikus sageduselt teisel kohal asuva sõna ja usalduspiirideks 3493 ± 118, s.o. 3375 ... 3611 (protsentides 3,4 ... 3,6). Nende kahe sõna usalduspiirid ei lõiku ja järelikult võib vähemalt 95 %-lise statistilise kindlusega väita, et sõna olema esineb eesti ilukirjandusproosa autorikõnes oluliselt sagedamini kui sõna ja. Sel juhul jäävad ka astakud muutumatuks (vastavalt 1. ja 2.). Võrreldes aga näiteks 5. ja 6. kohal asuvaid sõnu ei ning ei, mille sagedused on küllaltki lähedased (1395 ja 1300),

näeme, et usalduspiirid lõikuvad (1320 ... 1470 ja 1228 ... ... 1372). Järelikult ei saa nende sõnade astakute suhtes teha kindlat otsust, s.t. ei saa päris kindlasti öelda, et ei tingimata esineb üldkogumis sagedamini kui et.

Arvutades suhtelised vead, saame olema puhul 130/4237 = = 0,031 ehk 3,1 % ja ja puhul 118/3493 = 0,034 ehk 3,4 %. Sagedusjärjestuses 100.kohal asub lekseem sõna sagedusega F = 124. Esindusviga on sel juhul $2\sqrt{124} = 22$ ja usalduspiirid seega 102 ... 146 (protsentides 0,10 ... 0,15). Suhteline viga on siin tõusnud juba 17,7 %-ni (22/124). Mida väiksem sagedus, seda suurem on suhteline viga. Teatavasti võetakse sõnavarastatistilistes töödes kokkuleppeliselt "usaldusväärsuse" piiriks suhteline viga 30 %, mõnede uurijate arvates võib viga tõusta isegi kuni 45 %-ni (Бектаев, Пиотровский, 1974, 189).+ Nii arvestades võib usaldusväärseteks pidada tegelikult ainult 800–900 sõna meie lekseemide sagedussõnastikust. Kuid sellist traditsioonilist arvestusviisi ei saa sagedussõnastike puhul pidada kohaseks, sest sagedussõnastiku koostamisel ei seata eesmärgiks täpselt määrata iga sõna suhteline sagedus tekstis ning järjekorranumber sõnastikus, vaid olulisem on küllaldase tõenäosusega määrata sagedustsoon, kuhu kuulub üks või teine sõna. Näiteks võib kerkida küsimus ilukirjandusproosa autorikõne p õ h i s õ n a v a r a piiridest. Esitame omalt poolt järgmise kriteeriumi. Sõna loetakse allkeele põhisõnavarasse (s.o. kõrg- ja kesksagedustsooni) kuuluvaks, kui see esineb 100.000-sõnelises tekstis vähemalt 3 osavalimis 20-st (à 5000 sõnet) ning kogusagedusega vähemalt 10 ($F \geq 10$). Selliseid sõnu on ilukirjandusproosa autorikõne lekseemide sagedussõnastikus ligikaudu 1200 ja nende tekstikatmus on 75 % (vt. tabel 2). Sel alusel on koostatud lekseemide sagedussõnastiku lisas toodud täistähenduslike sõnade loetelud (vt. lk. 128 jj.). Sageduse F = 10 korral on usalduspiirid 4 ... 16. Võib kokku leppida selles, et põhisõnavara piiritsooni (üleminekutsooni) kuuluvad ka sõnad sagedusega 5 kuni 9 arvestusega, et F = 5 korral on usalduspiirid 1 ... 9, seega teoreetiline sageduse alampiir on üle 0. Sõnu sagedusega $F \geq 5$ on meie sagedussõnastikus kokku 2200 ja nad katavad tekstist umbes 80 % (vt. tabel 2). Ülejäänud sõnad ($F < 5$) tuleb pidada antud allkeele madalsagedusega sõnadeks, kusjuures ka nende hulgas võivad esineda antud allkeelele iseloomulikud sõnad, mida teeb kindlaks täiendav kvantitatiivne ja kvalitatiivne analüüs.

Ülaltoodud kriteeriumi aitaks täpsustada nn. stabiilsus- ja kasutuskoefitsientide arvutamine (vt. Kaasik, Tuldava, Villup, Käremaa, 1976, 109 jj.), mida aga lekseemide sagedussõnastiku puhul ei olnud võimalik tehnilistel põhjustel läbi viia. Sagedussõnastiku usaldatavust suudaks mõningal määral tõsta ka sõnastiku koostamisel kasutatud tekstivalimi suurendamine, kuid ökonoomsuse seisukohast annab see suhteliselt vähe tulu. Nii on näiteks kindlaks tehtud, et isegi 4,5 miljoni sõnelise valimi korral oleksid inglise keele sagedussõnastikus traditsioonilises mõttes "usaldusväärsed" (30 %-lise vea tasemel) ainult 1500 sõna (vt. Фрумкина, 1964, 13).

+ Foneetika- ja grammatikaalastes statistilistes uurimustes ei tohiks aga vea määr tõusta üle 20 % (vt. Бектаев, Пиотровский, 1974, 189).

Seos sõna sageduse ja astaku vahel (Zipfi seadus). Üks peamisi statistilisi seaduspärasusi, mis on kindlaks tehtud loomulike (traditsiooniliste, sotsiaalsete) keelte tekstide põhjal koostatud sagedussõnastike analüüsimisel, on funktsionaalse seos sõna (sõnavormi või lekseemi) sageduse ja astaku vahel, kusjuures astaku all mõeldakse sõna **järjekorranumbrit** sageduste kahanevas reas (vene k. ранг, saksa k. Rang, inglise k. rank). Esimesena märkas seda seaduspärasust prantslane J. Estoup (1916), kes tegeles stenograafilise kirja täiustamisega ja avastas, et sõnade paiknemine sagedussõnastikus allub erilistele statistilistele reeglitele. Ka füüsik E. Condon (1928) osutas sõnasageduste regulaarsusele ja püüdis leida matemaatilist väljendust sõna sageduse ja astaku vahekorrale. Funktsionaalse seose sõnade esinemissageduse ja järjekorranumbri vahel formuleeris lõplikult G.K. Zipf (1929). Nimelt leidis ta, et sõna absoluut- või suhteline sagedus ning vastav astak sagedussõnastikus on seotud küllalt pika teksti puhul järgmise sõltuvusega, mida on hakatud nimetama Z i p f i s e a d u s e k s:

$$F_r = \frac{C}{r^{\gamma}},$$

kus F_r tähistab sõna esinemissagedust antud tekstis, r - sõna astakut vastavas sagedussõnastikus, C ja γ on konstandid. Zipf tegi kindlaks, et konstandi γ väärtus kõigub tavaliselt 1 ümber. Juhul, kui $\gamma = 1$, võib seost väljendada: $F = \frac{C}{r}$ või $F_r \cdot r = C$, see tähendab, sageduse ja astaku korrutis peab andma mingi konstantse väärtuse (C).

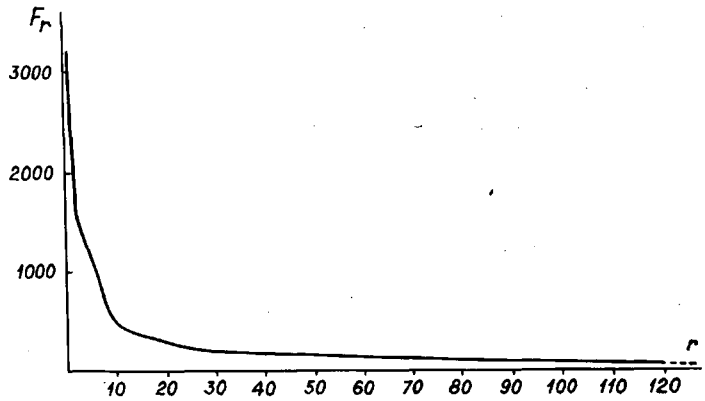
Võttes aluseks sõnade suhtelised sagedused, võime Zipfi seadust väljendada järgmiselt:

$$P_r = \frac{K}{r^{\gamma}},$$

kus P_r on sõna suhteline sagedus ja K on konstant, kusjuures $K = C/N$, kus N tähistab teksti pikkust sõnedes.

Seost sõnasageduse ja astaku vahel illustreerib joonis 2, millest on näha, et sagedussõnastiku alguses, s.t. kõige sagedamate sõnade puhul toimub sageduste langemine järsult, kuna aga vahemikus $10 < r < 20$ läheb langus üle mõõdukale kahanemisele. Antud juhul on meil tegemist andmetega eesti ilukirjandusproosa autorikõne sõnavormide sagedussõnastiku kohta (vt. tabel 1), kuid põhimõtteliselt samalaadiliselt toimub sageduste jaotumine kahanevas reas kõigi tuntud loomulike keelte puhul. See tähendab, et iga teksti puhul on ülemas väike rühm väga sagedaid sõnu, kusjuures nende hulgas on sageduste erinevused suhteliselt suured (järsk langus), ja suurem rühm keskmise ning väikese sagedusega sõnu, mis jaotuvad ühtlaselt kahaneva reana.

Zipfi valemi mõte seisneb selles, et tegelikkuses eksisteeriv diskreetne jaotus lähendatakse pideva jaotusega, mis annab meile joonisel 2 kujutatud hüperbooli taolise kõvera. Valemi konstandid tehakse kindlaks nn. vähimruutude meetodi abil (Tiit, 1972, 55 jj.). Kõige hõlpsamini saab arvutada valemi konstante lineariseerimise vahendusel. Antud juhul teeme seda võrrandi mõlema poole logaritmimisega.



Joon. 2. Seos sõnavormi sageduse (F_r) ja astaku (r) vahel ilukirjandusproosa autorikõne koondsagedussõnastiku andmete põhjal

Võttes arvesse, et Zipfi valemit saab esitada kujul $F_r = C \cdot r^{-\gamma}$, saame logaritmimisel: $\lg F_r = \lg C - \gamma \lg r$. Kui märkida $\lg F_r = Y$, $\lg C = A$, $\gamma = B$ ning $\lg r = X$, siis oleme saanud tavalise lineaarse regressioonivõrrandi: $Y = A + BX$.

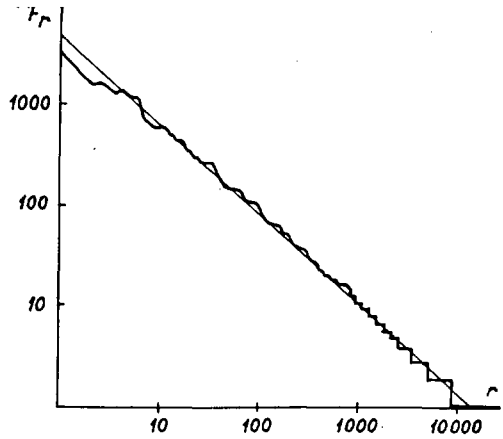
Kujutades ilukirjandusproosa autorikõne koondsagedussõnastikus esinevat seost sõnavormi sageduse ja astaku vahel täislogaritmdiagrammis, s.t. võttes mõlemad skaalad logaritmilised (vt. joon. 3), saame ühtlaselt langeva kõvera, mis on hästi lähendatav Zipfi valemi abil arvutatud tasandussirgele.

Tasandussirge arvutame vähimruutude meetodi abil. Tulemuseks saame lineaarse regressioonivõrrandi: $\lg F_r = 3{,}6151 - 0{,}8559 \lg r$. Logaritmide eemaldamisega saame Zipfi valemi

$$F_r = 4122 \cdot r^{-0{,}86} \qquad (\text{või} \quad P_r = 0{,}041 r^{-0{,}86}),$$

mis kirjeldab eesti ilukirjandusproosa autorikõne teksti sõnavormide tasandil. Konstandi γ väärtuseks on seega eestikeelse materjali põhjal 0,86.

Joonisest 3 nähtub, et madalamate sageduste puhul tekivad horisontaalsed platvormid, mille põhjuseks on asjaolu, et hulk sõnu esineb tekstis ühe ja sama sagedusega. Tasandussirge arvutamisel võtame aluseks vastava platvormi keskmise astaku, näiteks sagedusele F = 1, mis on omane sõnavormidele alates astakust 8974 kuni 30733 (vt. tabel 1), seame vastavusse keskmise astaku $r \approx 20.000$. Niisugune arvutusviis on täiesti õigustatud, arvestades seda, et tu-



Joon. 3. Seos sõnavormi sageduse ja astaku vahel. Täislogaritmdiagramm.

lemused peaksid olema ligilähedased sellele, mille saaksime, kui arvestaksime kõiki astakuid järjest. On juhitud tähelepanu asjaolule, et kõiki astakuid arvestades tekib vaid illusioon arvutuste täpsusest, sest tegelikult on sel juhul arvutus seotud süstemaatilise veaga (vt. Калинин, 1964, 125). Zipfi konstantide arvutuskäik eesti keele sõnavormide sagedussõnastiku põhal esitatakse täielikult tabelis 3.

Paljude keelte uurimisel on ilmnenud, et sagedussõnastiku alguses eraldub rühm kõrgsagedusega sõnu ("tuumik"), mille statistilised omadused erinevad teataval määral teiste sõnade omadustest. Täislogaritmdiagrammis moodustavad need sõnad tavaliselt omaette sirge, mis kaldub kõrvale üldisest tasandussirgest. Seda võib täheldada mõningal määral ka eestikeelse teksti puhul (vt. joon. 3), kusjuures tuumikusse kuuluvad umbes 10 kõige sagedamat sõna. Teiste keelte uurimisel on ilmnenud ka asjaolu, et sagedussõnastiku lõpposa erineb üldisest langustendentsist ja Zipfi tasandussirgest selle poolest, et madalsagedusega sõnu on vähem, kui neid peaks olema teoreetilise arvestuse järgi. Eesti keele sõnavormide sagedussõnastiku põhjal ei saa seda siiski väita: nagu nähtub joonisest 3, kirjeldab Zipfi tasandussirge hästi ka sagedussõnastiku lõpposa.

Et korvata Zipfi valemi puudust, nimelt mittetäielikku vastavust sagedussõnastiku algosa (tuumiku) kirjeldamisel, kasutatakse tavaliselt tasandussirge arvutamisel nn. M a n d e l b r o t i p a r a n d u s t. Prantslane B. Mandelbrot esitas oma paranduse seoses teoreetiliste kaalutlustega teksti genereerimise kohta (Mandelbrot, 1954). Parandatud valem, mida kutsutakse Zipfi-Mandelbroti seaduseks, on järgmine:

$$F_r = \frac{C}{(r+\varrho)^{\delta}} \qquad (\text{või: } P_r = \frac{K}{(r+\varrho)^{\delta}}),$$

kus ϱ tähistab Mandelbroti parandusliiget.

Zipfi-Mandelbroti valem annab hea vastavuse empiiriliste andmetega sagedussõnastiku algstaadiumis ja alates $r > 15$ ei mõjuta praktiliselt kõvera kuju. Eesti keele materjali põhjal on q suurus väike, sest tuumiku hälbimine üldisest tendentsist pole kuigi suur. Iteratsioonimeetodi abil saame q väärtuseks 0,5, kusjuures konstandid C ja γ suurenevad mõningal määral.

Zipfi-Mandelbroti valem eesti keele sõnavormide sagedussõnastiku põhjal on järgmine:

$$F_r = \frac{4585}{(r + 0{,}5)^{0{,}87}}$$

või suhteliste sageduste puhul

$$P_r = \frac{0{,}046}{(r + 0{,}5)^{0{,}87}} .$$

Kontrolliks võib küsida, kui suur peaks olema näiteks sagedusjärjestuses 580. kohal asuva sõnavormi sagedus ilukirjandusproosa autorikõne sagedussõnastikus. Arvutame Zipfi-Mandelbroti valemi järgi:

$$P_{580} = 0{,}046\ (580 + 0{,}5)^{-0{,}87} = 0{,}000181$$

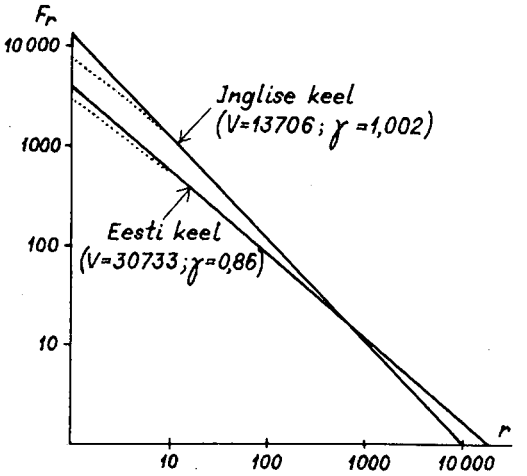
(s.o. 0,0181 %).

Et koondvalimi maht on 100.000 (täpsemalt N = 99898), siis saame sõnavormi absoluutsageduseks P·N = 18,1. Sõnavormide sagedussõnastikus on ka tegelikult 580. sõnavormi sageduseks F = 18. Muide ka Zipfi valemiga arvutades (ilma Mandelbroti parandusliikmeta) saame ligilähedase tulemuse, nimelt F = 17,3.

Eelnevast selgub, et Zipfi valemi parameetride tundmine teatud keele või allkeele ülatuses omab praktilist tähtsust. Valemi abil saab prognoosida sagedussõnastiku mahtu ja struktuuri, mis on eriti tähtis tekstide automaattöötlemisel (näiteks on võimalik ratsionaalsemalt jaotada masinamälu). Peale selle on Zipfi seaduse alusel võimalik ette arvestada valimi mahtu vajaliku representatiivsuse tagamiseks. Huvi pakub ka asjaolu, et valemi parameetrid kujutavad endast stiile eristavaid karakteristikuid ja eri keelte võrdlemisel saab nende abil teha keeletüpoloogilisi järeldusi. Eesti ilukirjandusproosa autorikõne põhjal saime Zipfi valemi konstandi γ väärtuseks 0,86. Niisama suure teksti põhjal arvutatud γ inglise keele kohta on 1,002 (Kučera, Francis, 1967, 357). Joonis 4 illustreerib eesti ja inglise sõnavara struktuuri erinevust Zipfi tasandussirge abil. Nagu näha, on inglise keele puhul sirge tõusunurk suurem (mida väljendab γ kui tõusunurga tangens), see tähendab, et kõrgsagedusega sõnavormid katavad hulga suurema osa tekstist kui eesti keeles ning sõnavormide tagavara lõpeb inglise keeles kiiremini. Erinevus keelte vahel tuleneb esmajoones sellest, et eesti keeles on palju muutevorme. Seetõttu on ka eri sõnavormide arv eesti keele 100.000-sõnelises tekstis üle 30.000, kuna aga inglise keele vastavas tekstis esineb ainult ligikaudu 14.000 sõnavormi.

Seni vaatlesime Zipfi seaduse toimet sõnavormide tasandil, Põhimõtteliselt sama laadi on sagedussõnastiku

struktuur ka lekseemide tasandil, ainult parameetrite väärtused muutuvad. Eesti ilukirjandusproosa autorikõne l e k s e e m i d e sagedussõnastiku põhjal arvutatud Zipfi valemi konstandid on järgmised: C = 6823, K = 0,068, γ = 0, 92. Tõusunurk täislogaritmdiagrammis on seega lekseemide puhul suurem kui sõnavormide tasandil.



Joon. 4. Zipfi seadus eesti ja inglise keele tekstide põhjal sõnavormide tasandil (valimid à 100.000 sõnet; V tähistab sõnastikku mahtu). Täislogaritmdiagramm.

Ühe ja sama keele tekstide puhul peab arvestama, et γ väärtus oleneb suurel määral teksti pikkusest. Nii näiteks saame ilukirjandusproosa autorikõne sagedussõnastikus osalenud üksikvalimi (teksti pikkus N = 5000) puhul Zipfi valemi parameetrite väärtusteks:

sõnavormide tasandil C = 129, K = 0,026 ja γ = 0,68;

lekseemide tasandil C = 196, K = 0,039 ja γ = 0,72.

Zipfi seadus väljendab teatavat universaalset omadust, mis puudutab kõiki traditsioonilisi keeli. See omadus seisneb selles, et igasuguse teksti (kõne) peamise osa moodustab väike arv sageli korduvaid sõnu, kuna aga ülejäänud osa koosneb tuhandetest sõnadest, mis esinevad harva. Iseloomulik igasugusele tekstile on see, et kasutatud sõnadest saab koostada sagedussõnastiku, mille järjestatud sagedused moodustavad kahaneva rea Zipfi valemi alusel. See on küllaltki omapärane nähtus, mida on proovitud seletada nii psühholoogiliste kui ka muude põhjustega. Zipf ise arvas, et kõne elementide hierarhiline organisatsioon ("harmooniline rida") on seotud "minimaalse jõupingutuse" printsiibiga. Psühholingvistiline seletus põhineb teesil, et "kõigil keele tasanditel konkureerivad alternatiivid, mis on organiseeritud hierarhiliselt esinemissageduse alusel

ning on suhteliselt madala jaotus-entroopiaga" (Osgood, 1966,304). See tähendab seda, et kui inimene peab korduvalt valima alternatiivide seast, siis valib ta väiksema hulga alternatiive sageli ja suurema hulga väga harva. Arvatakse, et selline hierarhiline organisatsioon ei puuduta mitte ainult inimkeelt, vaid et igasugune elav organism on oma loomult entroopiline (vt. Osgood, 1966, 205). Sõnasageduste reastumine ligikaudselt Zipfi kõvera järgi peaks väljendama koodi optimaalsust, mis on ajalooliselt välja kujunenud keele arengu käigus. B. Mandelbrot (1954) seletab Zipfi kõvera tekkimist ökonoomiaga teadete edastamisel sõnade abil, see ökonoomia seisneb selles, sõnade informatsiooniline sisu (mida määrab nende sagedus) on proportsionaalne sõnade keerukusega (mida määrab nende astak). Nõukogude teadlase J. Šreideri (Шрейдер, 1967) arvates on sõna keerukus esmajoones seotud sõna tähendusega, sel juhul saab paralleele tõmmata "tähendusjärgu" ja "sagedusjärgu" vahele. Praktiliselt tähendab see seda, et näit. sagedamatel sõnadel on suurem tähenduste spekter. Väärib tähelepanu ka J. Orlovi teooria, mis rõhutab Zipfi seaduse paikapidavust esmajoones terviklike ja kunstiliselt kõrge väärtusega tekstide puhul (Орлов, 1969 ja 1970). See kinnitab teatud määral arvamust, et funktsionaalne seos sõna sageduse ja astaku vahel on seotud optimaalse koodiga. Sel juhul on ka Zipfi funktsiooni parameeter γ lähedane 1-le, mis tegelikult leiab aset suuremas osas loomuliku keele tekstides (eriti lekseemide vasandil). Uurijad on juhtinud tähelepanu sellele, et kui γ väärtus oluliselt erineb 1-st, siis võib olla tegemist patoloogiliste juhtumitega (Фрумкина, 1964, 27). Ka lastekeeles on γ väärtus erinev tavalisest, mis on seletatav väiksema sõnavaraga (Mandelbroti eksperimendi andmeil oli lastekeeles γ väärtuseks 1,6, kusjuures see väärtus kahanes vähehaaval kuni 1-ni lapse täiskasvanuks saades, vt. Pierce, 1962, ptk. 12).

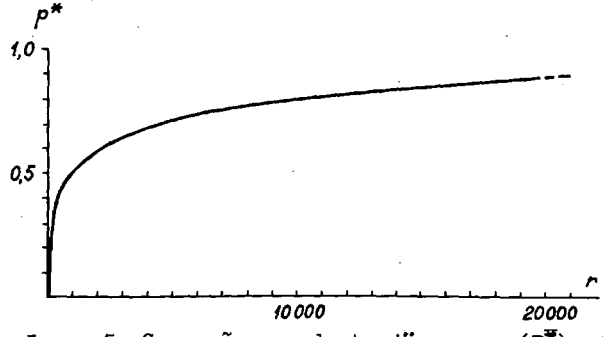
Zipfi-Mandelbroti funktsiooni kasutatakse viimasel ajal mitte ainult loomulike keelte, vaid ka nn. informatsioonikeelte statistilisel uurimisel. On selgunud, et deskriptorite jaotus allub tavaliselt Zipfi seadusele (Ваха-бов, 1970). See seik etendab tähtsat osa informaatika probleemide lahendamisel, millega seletub ka suur huvi, mis viimasel ajal valitseb Zipfi seaduse vastu informaatikaalaste teaduslike ajakirjade veergudel. On kindlaks tehtud, et Zipfi seadusel on laiem rakendusSfäär, kui seda on keelelised küsimused. Nii näiteks võrdub Zipfi seadus Bradfordi seadusega informatsiooni jaotuse kohta ajakirjanduses, Lotka seadusega teadlaste produktiivsuse kohta jne. (Шрейдер, 1967; Козачков, 1969). Zipfi seadus kehtib isegi sellistel juhtudel nagu linnade suuruse jaotus ühe maa piirides (Pierce, 1962, ptk. 12). Antud juhul on tegemist kokkusattuvusega, mis iseloomustab nähtusi kui ajalooliselt kujunenud ökonoomseid (optimaalseid) struktuure.

Võib veel lisada, et seost sõna sageduse ja astaku vahel saab vaadelda ka lähtudes sageduste järgsummadest (kumulatiivsest sagedusest). Astaku (r) suurenedes kasvab ka sageduste järgsumma (F^{*} - absoluutsageduste, P^{*} - suhteliste sageduste korral), kuid selliselt, et alates teatud momendist vastab suurele r kasvule väike F^{*} (või P^{*}) kasv. Seda illustreerib joonis 5. Sisuliselt on siin tegemist teksti katmusega. Sel juhul kirjeldab sõna sageduse

ja astaku vahelist seost nn. Weibulli funktsioon, mida keelematerjali põhjal kasutas esmakordselt nõukogude teadlane G. Belonogov (Белоногов, 1962). Funktsioon leitakse väljendi abil:

$$P^{*} = 1 - e^{-cr^{k}},$$

kus P^{*} tähistab suhteliste sageduste järgsummat, r - astakut, e on naturaallogaritmide alus (e = 2,7183) ja c ning



Joon. 5. Seos sõnasageduste järgsumma (P^{*}) ja astaku (r) vahel

k on konstandid (parameetrid). Arvutades eesti keele ilukirjandusproosa autorikõne sõnavormide sagedussõnastiku põhjal funktsiooni parameetrid, saame järgmised tulemused, mida kõrvutame andmetega teiste keelte kohta (vt. Водисевич, 1972, 8):

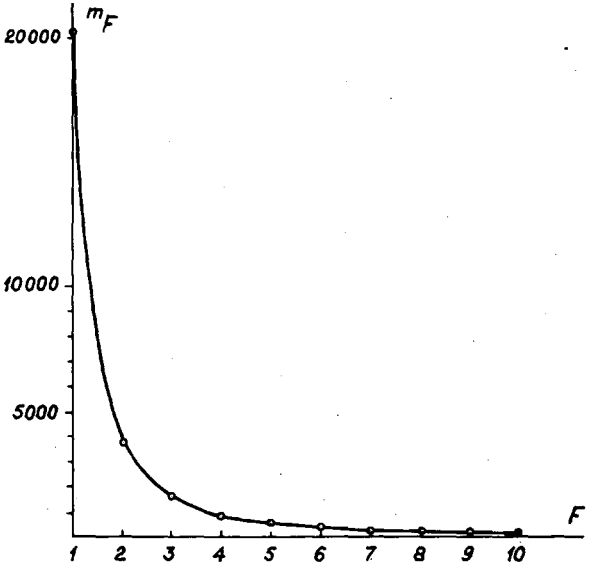
	c	k
eesti keel	0,053	0,379
prantsuse keel	0,143	0,357
inglise keel	0,155	0,330

Võib nentida olulist erinevust sünteetilise eesti keele ja analüütiliste prantsuse ning inglise keele vahel eriti parameetri c väärtuses. Weibulli-Belonogovi valem väljendab tegelikult samu suhteid mis Zipfi seadus, kusjuures parameetrid c ja C on lähedased, kuna aga k ja γ korreleeruvad negatiivselt. Weibulli-Belonogovi funktsiooni paremuseks on see, et sagedussõnastiku lõpposas on valemi vastavus parem kui Zipfi valemi järgi arvutades. Traditsiooniliselt on aga Zipfi valem enam kasutatav.

Leksikaalne spekter (Yule'i jaotus). Teksti ja sõnastiku statistilist struktuuri saab kirjeldada ka sel teel, et tekstis esinevad sõnad rühmitatakse esinemissageduse järgi gruppidesse ja tehakse kindlaks gruppi kuuluvate sõnade esinemisarv. Nende andmete alusel koostatakse tabel, milles seatakse vastavusse sõna (sõnavormi või lekseemi) sagedus F ja esinemisarv m_F. Selline tabel väljendab antud teksti sõnavara sagedusjaotust ehk "leksikaalset spektrit". Tavaliselt moodustatakse leksikaalse spektri laiendatud tabelid, milles absoluutarvude kõrval esinevad ka suhtarvud (suhtelised sagedused protsentides või kümnendmurdudes) ja kumulatiivsed sagedused ehk järgsummad. Sel juhul kõneldakse Yule'i jaotusest. Tabelis võib välja arvutada ka teksti

katmuse nii absoluut- kui ka suhtarvudes. Kokkuvõtlikud andmed eesti ilukirjandusproosa autorikõne sõnavormide ja lekseemide sagedussõnastike põhjal esitame tabelites 4 ja 5.

Empiiriliselt on kindlaks tehtud fakt, et olenemata teksti iseloomust ja pikkusest võib igas küllalt pikas tekstis konstateerida pöördvõrdelist suhet esinemissageduse (F) ja vastava esinemisarvu (m_F) vahel, s.t. mida väiksem on sagedus, seda rohkem on sõnu selle sagedusega, ja vastupidi, mida suurem on sagedus, seda vähem esineb selle sagedusega sõnu. Kehtib kindel seaduspärasus, mille kohaselt üks kord esinevate sõnade (s.o. sagedusega F = 1) hulk on igasuguses normaalses tekstis kõige suurem, kaks korda esinevaid sõnu on rohkem kui kolm korda esinevaid sõnu jne. See seaduspärasus on universaalne ja kehtib kõigi loomulike keelte suhtes.



Joon. 6. Sõnavormide sagedusjaotus (Yule'i kõver) koondsagedussõnastiku andmete põhjal

Esimesena uuris niisugust sõnade jaotust matemaatiliselt inglane G.U. Yule (1944), kes illustreeris jaotust graafikul (meie materjali põhjal vt. joon. 6) ja juhtis tähelepanu järgmistele iseärasustele: esiteks - esinemissageduse ja esinemisarvu vahelist funktsionaalset seost väljendaval kõveral on olemas ainult üks maksimum (kõrgeim punkt), mis vastab esinemissagedusele F = 1; teiseks - F > 1 korral toimub algul järsk langus, mis vähehaaval läheb üle mõõdukale kahanemisele; kolmandaks - kõige suurema F suuruse puhul on esinemisarv alati 1, s.t. igas tekstis saab olla ainult üks kõige sagedam sõna (sõnavorm resp. lekseem). Kõverat, mis näitab esinemissageduse (F) ja esinemisarvu (m_F)

vahelist seost, nimetatakse keelestatistikas Y u l e ' i k õ v e r a k s.

Nagu nimetasime, väljendab Yule'i kõver kõigi loomulike keelte tekstide statistilist struktuuri, kuid seda tuleb mõista kui sarnasust üldjoontes. Detailsem vaatlus võimaldab kindlaks teha olulisi erinevusi sõnasageduste jaotumuses, mis peegeldub ka kõvera kujus. Näiteks võib ühes tekstis olla väikeste sagedustega sõnu suhteliselt rohkem, teises tekstis aga vähem; samal ajal võib aga esimene tekst sisaldada ka rohkem suurte sagedustega sõnu jne. Erinevaid kombinatsioone võib olla mitmesuguseid. See lubab kasutada Yule'i kõverat ja vastavat sagedusjaotust (leksikaalset spektrit) tekstide stilostatistilisel võrdlemisel (näit. vt. Tuldava, 1971, 234 jj.). Yule'i jaotus toob esile ka tüpoloogilised erinevused keelte kõrvutamisel. Eri keelte tekste võib vaadelda sellest seisukohast, mil määral teatud esinemissagedusega sõnad katavad vastavat sõnastikku ja teksti. Võrreldavad tekstid peavad olema ühesuurused, sest Yule'i jaotuse korral kehtib seaduspärasus, et teksti pikkuse suurenedes väheneb üks kord esinevate sõnade osatähtsus sõnastikus ja sellest tulenevalt toimuvad ka teised muutused kogu sõnavara struktuuris.

Vaatleme võrdlevalt andmeid eesti ilukirjandusproosa autorikõne, vene keele elektroonikatekstide (Калинина, 1968, 104-105) ja inglise keele segatekstide kohta (Kučera, Francis, 1967, 327-329), kusjuures kõik valimid on ühesuurused (N = 100.000) ja sõnastikke vaadeldakse sõnavormide tasandil:

Sõnavormi esinemissagedus (F)	Sõnastiku katmus (%) Eesti keel	Vene keel	Inglise keel	Teksti katmus (%) Eesti keel	Vene keel	Inglise keel
1	70,8	45,3	51,6	21,8	6,4	7,0
1 - 5	93,1	79,8	83,5	40,3	20,3	19,4
1 - 10	96,6	89,3	91,5	48,5	30,5	27,6
1 - 100	99,7	99,3	99,4	73,0	67,7	53,2

Kõigepealt võib nentida üldist seaduspärasust, mille järgi üks kord esinevate sõnade arv moodustab kõigis keeltes suure osa sõnastikust (45-70 %), kuid teksti katmus nende sõnadega on märksa väiksem (6-22 %). Tabelist võib välja lugeda ka seda, et väga sagedad sõnad (sagedusega üle 100) hõlmavad sõnastikust ainult 0,3-0,7 %, kuid tekstist katavad nad 27-47 %. Üldiste ühisjoonte kõrval võib täheldada olulisi lahkuminekuid eri keelte sagedusjaotustes. Sõnastikust katavad üks kord esinevad sõnavormid ja üldse madala sagedusega sõnavormid (näit. F = 1 ÷ 5) eesti keele sõnastikus ja tekstis märgatavalt suurema osa kui vene ja inglise keeles. See on seletatav suure arvu muutevormidega, mis iseloomustavad eesti keelt. Vene ja inglise keele võrdlemine osutab aga antud juhul küllaltki suurt lähedust, mis ei ole päris ootuspärane. Ka vene keel kui sünteetilisem keel peaks näitama suuremat sõnastiku ja teksti katmust kui inglise keel. Käesoleval juhul on aga tegemist vene keele tehnika-alaste tekstidega, mis oma sõnavaralt on vaesemad kui vastavad ilukirjanduslikud või segatekstid. Sagedusjaotuse tüüp on seoses ka sõnavara rikkusega, seepärast peaks eri keeli võrdlema ühtede ja samade või lähedaste žanride valdkonnas. Meie näitega oli võimalus

demonstreerida erinevate keelte ühiseid jooni, vaatamata žanride erinevustele, ja lahkuminekuid žanride lähedase korral. Erinevus eesti ja inglise keele struktuuri vahel (lähedaste žanride puhul) torkab silma eriti teksti tasandil, näit. üks kord esinevad sõnavormid katavad eesti keeles 22 %, inglise keeles aga ainult 7 % tekstist.

Teoreetilisest ja praktilisest seisukohast pakub huvi sõnasageduste (F) ja esinemisarvu (m_F) vahelise seose matemaatiline väljendus. On tehtud hulgliselt katseid leida sobivat matemaatilist mudelit nimetatud seose kirjeldamiseks. Nõukogude teadlased V. Kalinin (Калинин, 1964 ja 1965), J. Orlov (Орлов, 1969 ja 1970) ning I. Nadareišvili (Надареишвили, Орлов, 1971) seostavad leksikaalse spektri Zipfi seadusega ja esitavad vastavad valemid, mille abil saab arvutada eri sagedusega sõnade teoreetilisi esinemisarve tingimusel, et Zipfi seadus kehtib teksti kogu ulatuses või osaliselt. Uuematest uurimustest väärib tähelepanu M. Arapovi ja E. Jefimova (Арапов, Ефимова, 1975) käsitlus teksti leksikaalsest struktuurist, mille alusel saab arvutada leksikaalse spektri väärtused, toetudes "diskreetsuse" faktoriga parandatud Zipfi seadusele.

Matemaatilise mudeli otsimise mõte seisneb selles, et antud juhul saaks prognoosida eri sagedusega sõnade esinemisarve, lähtudes vaid osalisest informatsioonist teksti statistilise struktuuri kohta. Sellistel prognoosidel on suur praktiline tähtsus tekstide automaattöötlemisel ja ka stilostatistiliste probleemide lahendamisel. Puhtteoreetiliselt tuletatud mudelite kõrval on esitatud ka rida empiirilisi valemeid, mis kehtivad antud keele või allkeele ulatuses (näit. Simon, 1960; Carroll, 1967).

Universaalsema iseloomuga on G. Herdani (1964) valem, mis põhineb 18. sajandi inglise matemaatiku E. Waringi poolt uuritud jaotusel. Keelestatistikas on hakatud seda nimetama W a r i n g i - H e r d a n i s e a d u s e k s. Valemi järgi saab leida teoreetilised sõnasageduste esinemisarvud, kui on teada teksti maht (N), sõnastiku maht (V - sõnavormide või L - lekseemide tasandil) ning üks kord esinevate sõnavormide (V_1) või lekseemide (L_1) arv. Waringi jaotus kujutab endast kahanevate tõenäosuste rida, mida määravad parameetrid x ja a:

$$\frac{x-a}{x} + \frac{(x-a)a}{x(x+1)} + \frac{(x-a)\ a\ (a+1)}{x(x+1)\quad (x+2)} + \ldots$$

$$\ldots + \frac{(x-a)\ a\ (a+1)\ \ldots\ (a+n-1)}{x(x+1)\ (x+2)\ldots\ (x+n)} = 1.$$

Parameetrid a ja x leitakse järgmiselt:

$$a = \frac{1}{\frac{1}{Q} - T - 1}, \qquad x = \frac{a}{Q},$$

kusjuures $Q = 1 - (V_1/V)$ või $Q = 1 - (L_1/L)$ ja $T = V/N$ (või L/N).

Antud sagedusega sõnade esinemise tõenäosus (p_i) leitakse Waringi jaotuse tõenäosuste reast vastavas järjekor-

ras, s.t. üks kord esinevate sõnade esinemise tõenäosus $p_1 = \frac{x-a}{x}$, kaks korda esinevate sõnade esinemise tõenäosus $p_2 = \frac{(x-a)a}{x(x+1)}$ jne. Korrutades tõenäosused iga kord sõnastiku mahuga (V või L), saame vastavad absoluutarvud - sõnasageduste esinemisarvud.

Kontrollime Waringi-Herdani seaduse paikapidavust eesti keele materjali põhjal. Vaatleme kõigepealt ilukirjandusproosa autorikõne koondvalimit sõnavormide tasandil. Lähteandmed on järgmised: N = 99898, V = 30733 ja V_1 = = 21760(vt. tabel 4). Arvutame vahepealsed väärtused: Q = = 1 - V_1/V = 0,292 ja 1/Q = 3,425; T = V/N = 0,218; a = = 0,453 ja x = 1,551. Üks kord esinevate sõnavormide esinemise tõenäosus $p_1 = \frac{x-a}{x}$ = 0,708 (s.o. 70,8 %; see vastab täpselt empiirilisele väärtusele, vt. tabel 2). Kaks korda esinevate sõnavormide tõenäosus $p_2 = (\frac{x-a}{x})(\frac{a}{x+1})$ = = 0,708 $(\frac{0,453}{2,551})$ = 0,126; kolm korda esinevate sõnavormide tõenäosus $p_3 = \left[\frac{(x-a)a}{(x+1)}\right]\left[\frac{a+1}{x+2}\right]$ = 0,126 $(\frac{1,453}{3,551})$ = = 0,052 jne. Korrutades tõenäosused sõnastiku mahuga (V), saame iga kord vastava sagedusega sõnavormide esinemisarvu. Saadud teoreetilist sõnasageduste jaotust võrdleme empiirilise jaotusega ja kontrollime homogeensuse hüpoteesi hii-ruut-testi abil (vt. tabel 6). Tulemus (χ^2 = 28,3) on väiksem kriitilisest hii-ruudu väärtusest 5%-lisel olulisusnivool ja järelikult võib väita, et erinevused empiirilise ja teoreetilise jaotuse vahel on juhuslikku laadi. Kontrollimine näitab, et ka väiksema valimi korral (N = = 5000) on vastavus empiirilise ja Waringi-Herdani valemi abil arvutatud teoreetilise jaotuse vahel hea. Seega võib öelda, et Waringi-Herdani seadus kehtib eestikeelse teksti puhul sõnavormide tasandil (esialgu võib seda väita tekstide kohta mahuga 5000-100.000 sõnet). Lekseemide tasandil (koondvalimi lähteandmed: N = 99898, L = 14654, L_1 = 8682; vt. tabel 5) on aga erinevused empiiriliste ja teoreetiliste andmete vahel suuremad, eriti väiksemate sageduste puhul (F = 2 ja F = 3). Leksikaalse spektri algusosa kohta tehtud kontroll näitab, et jaotuste erinevus on statistiliselt oluline, kuid ligikaudse hinnangu järgi (vt. tabel 7) võib Waringi-Herdani jaotust pidada küllaltki lähedaseks eesti sõnasageduste jaotumusele lekseemide tasandil.

\+ + +

Artiklis analüüsiti eesti keele sõnavara ja teksti kvantitatiivset struktuuri, lähtudes tänapäeva ilukirjandusproosa autorikõne sagedussõnastiku andmetest. Olemasoleva kvantitatiivse tüpoloogia skeemi järgi kuulub eesti keel "sünteetilis-flektiivsete" keelte hulka. Eestikeelsele tekstile on omane sõnavormide vähene kontsentratsioon tekstis, nii näit. kordub 100.000-sõnelises tekstis iga sõnavorm keskmiselt 3,25 korda (niisama suures inglise keele tekstis on sõnavormi korduvus keskmiselt 7,1). 100

sagedamat sõnavormi katavad eestikeelsest ilukirjanduslikust tekstist 28–30 % (inglise keeles 48–50 %). Üks kord esinevaid sõnavorme on eestikeelses tekstis 22 % ja vastavas sõnastikus 71 % (inglise keeles vastavalt 7 ja 52 %). Ilukirjandusproosa autorikõne sagedussõnastiku põhjal võis täheldada teksti head vastavust nn. Zipfi seadusele (Zipfi konstandi γ väärtused on sõnavormide tasandil 0,86 ja lekseemide tasandil 0,92). Eestikeelse teksti leksikaalset spektrit (sõnasageduste jaotumust) kirjeldab sõnavormide tasandil eriti hästi Waringi-Herdani funktsioon.

Nimetatud empiirilised seaduspärasused võimaldavad teha praktilisi järeldusi keelestatistiliseks uurimistööks. Sagedussõnastiku põhjal tehtav uurimistöö ei piirdu puhtkvantitatiivse vaatlusega, vaid objektiivsete statistiliste andmete alusel avanevad võimalused mitmekülgseks kvalitatiivseks analüüsiks nii lingvistilis-stilistiliste kui ka pedagoogilis-psühholoogiliste probleemide lahendamisel.

Viidatud kirjandus

Allén, S. Nusvensk frekvensordbok, baserad på tidningstext. Stockholm, Almqvist & Wiksell, 1970.

Allén, S. Vocabulary Data Processing. - The Nordic Languages and Modern Linguistics. Proceedings of the International Conference of Nordic and General Linguistics. Reykjavík, 1970, 235–261.

Bečka, J. V., Jelínek, J., Těšitelová, M. Frekvence slov, slovních druhů a tvarů v českém jazyce. Praha, 1961.

Carroll, J. B. On Sampling from a Lognormal Model of Word-Frequency Distribution. - Computational Analysis of Present-Day American English. By H. Kučera and W.N. Francis. Providence, 1967, 406–424.

Condon, E. U. Statistics of Vocabulary. - "Science", vol.67, 1928.

Eaton, H. S. An English-French-German-Spanish Word Frequency Dictionary. A Correlation of the First Six Thousand Words in Four Single-Language Frequency Lists. New York, 1961.

Engwall, G. Fréquence et distribution du vocabulaire dans un choix de romans français. Stockholm, Skriptor, 1974.

Estoup, J. B. Gammes sténographiques. Ed. 4. Paris, 1916.

García Hoz, V. Vocabulario usual, común y fundamental. Madrid, 1953.

Hauch, E. A German Idiom List Selected on the Basis of Frequency and Range of Occurrence. New York, 1931.

Herdan, G. Quantitative Linguistics. London, 1964.

Hinderling, R. Rückläufiges Estnisches Wörterbuch, I. Das Material der Grundformen. Regensburg, 1975.

Josselson, H. The Russian Word Account and Frequency Analysis of Grammatical Categories of Standard Literary Russian. Detroit, 1953.

Juilland, A., Brodin, D., Davidovitch, C. Frequency Dictionary of French Words. The Hague, Mouton, 1970.

Kaasik, U., Tuldava, J., Villup, A., Ääremaa, K. Eesti ilukirjandusproosa autorikõne sõnavormide sagedussõnastik. - Töid keelestatistika alalt, 1. TRÜ Toimetised, vihik 377. Tartu, 1976, 107–153.

Kaeding, F. Häufigkeitswörterbuch der deutschen Sprache. Steglitz bei Berlin, 1898.

Kučera, H., Francis, W. N. Computational Analysis of Present-Day American English. Providence, R. I., 1967.

Mandelbrot, B. Structure formelle des textes et communication.- "Word", vol. 10, 1954, No. 1, 1-27.

Mistrík, J. Frekvencia slov v slovenčine. Bratislava, 1969.

Osgood, Ch. E. Language Universals and Psycholinguistics. - Universals of Language. Cambridge, Mass., 1966.
Papp, F. Die Untersuchung des Wortschatzes und der Morphologie der Sprachdenkmäler mit Hilfe der Konkordanz. - Congressus Quartus Internationalis Fenno-Ugristarum. Tézisek. Budapest, 1975, 31.
Pierce, J. R. Symbols, Signals and Noise: the Nature and Process of Communication. London, Hutchinson, 1962.
Saukkonen, P. Statistical Viewpoints in Stylistics. - Congressus Quartus Internationalis Fenno-Ugristarum. Tézisek. Budapest, 1975.
Simon, H. A. Some Further Notes on a Class Skew Distribution Functions. - "Information and Control", 1960, 3, 80-81.
Słownik polszczyzny XVI w. Zeszyt próbny. Wrocław, 1956.
Štindlová, J. Les dictionnaires inverses. - "Cahiers de lexicologie", vol. 2. Paris, 1960, 79-86.
Zipf, G. K. Relative Frequency as a Determinant of Phonetic Change. - Harvard Studies in Classical Philology. No. 40. Cambridge, Mass., 1929.
Tauli, V. Word Index to August Mälk's Tee kaevule I. The Institute of Finno-Ugric Languages. Uppsala, 1964.
Thorndike, E. L., Lorge, J. The Teacher's Word Book of 30,000 Words. New York, Columbia University, 1944.
Tiit, E. Matemaatilise statistika tabelid, II. Tartu, TRU, 1972.
Tuldava, J. Sõnavara statistilisest struktuurist. - Linguistica, III. Tartu, TRU, 1971, 211-248.
Tuldava, J. Eesti keele sõnavara statistiline struktuur. Käsikiri (säilitatakse TRU saksa filoloogia kateedris). Tartu, 1976.
Tuldava, J. Statistilised jaotused keeleteaduses. -Linguistica, VIII. TRU Toimetised, vihik 401. Tartu, 1976a, 111-125.
West, M. A General Service List of English Words with Semantic Frequencies and a Supplementary Word List for the Writing of Popular Science and Technology. London, 1953.
Õigekeelsuse sõnaraamat. Toimetanud E. Nurm, E. Raiet ja M. Kindlam. Tallinn, ERK, 1960.
Yule, G. U. The Statistical Study of Literary Vocabulary. Cambridge University Press, 1944.

Алексеев П. М. Частотный словарь английского подъязыка электроники. Автореф. канд. дисс. Л., 1965.
Алексеев П. М., Турыгина Л. А. Частотный англо-русский словарь-минимум газетной лексики. М., 1974.
Арапов М. В., Ефимова Е. Н. Понятие лексической структуры текста. - Научно-техническая информация. Серия 2. М., 1975, № 6, 3-7.
Бектаев К. Б. Статистика речи 1957-1972. (Библиографический указатель.) Алма-Ата, 1972.
Бектаев К. Б. Статистико-информационная типология тюркского текста. Автореф. докт. дисс. Л., 1975.
Бектаев К. Б., Пиотровский Р. Г. Математические методы в языкознании. Часть 2. Математическая статистика и моделирование текста. Алма-Ата, 1974.
Белоногов Г. Г. О некоторых статистических закономерностях в русской письменной речи. - "Вопросы языкознания", 1961, №1.
Вибергштейн Б. Ш. Автоматическое построение словарей-конкордансов. - Вопросы лингвостатистики и автоматизация лингвистических работ. Вып. 5. Труды ЦНИИПИ, сер. 3. М. 1971, 42-43.
Борисевич А. Д. Англо-русский автоматический словарь оборотов. Автореф. канд. дисс. Минск, 1972.
Вахабов В. К. Некоторые результаты статистических исследований дескрипторного языка. - Научно-техническая информация. Серия 2. М., 1970, № 4, 27-31.

Вашак П. Длина слова и длина предложения в текстах одного автора. - Вопросы статистической стилистики. Киев,"Наукова думка", 1974, 314-329.
Вертель В. А., Вертель Е. В., Крисевич В. С., Пиотровский Р.Г., Трибис Л. И. Автоматические словари в системе бинарного вероятностного МП. - Инженерная лингвистика. Учен. зап. ЛГУ, том 458, часть 1, 1971, 4-147.
Данейко М. В., Машкина Л. Е., Нехай О. А., Соркина В. А., Шаранда А. Н. Статистическое исследование лексической дистрибуции словоформы. - Статистика речи. Л., "Наука", 1968,
Засорина Л. Н. Частотный словарь современного русского языка. М., "Советская энциклопедия", 1977.
Зореф М. Г. Частотный словарь немецких текстов по электронике. - Статистика речи и автоматический анализ текста. Л., "Наука", 1971, 229-252.
Калинин В. М. О статистике литературного текста. - "Вопросы языкознания", 1964, № 1, 123-127.
Калинин В. М. Функционалы, связанные с распределением Пуассона, и статистическая структура текста. - Труды Математического Института им. Стеклова. Том 29. М., 1965, 182-197.
Калинина Е. А. Изучение лексико-статистических закономерностей на основе вероятностной модели. - Статистика речи. Л.,"Наука", 1968, 64-107.
Ковачков Л. С. Некоторые методологические вопросы теории информационно-поисковых систем. - Научно-техническая информация. Серия 2. М., 1969, № 12, 9-16.
Надарейшвили И. Ш., Орлов Ю. К. Рост лексики как функция длины текста. - Сообщения АН Грузинской ССР. Том 64, № 3, 1971, 549-552.
Орлов Ю. К. Обобщение закона Ципфа. - Тезисы межвузовской конференции "Проблемы прикладной лингвистики", 16-19 декабря 1969 г. Часть 2. М. 1969.
Орлов Ю. К. О статистической структуре сообщений оптимальных для человеческого восприятия (к постановке вопроса). - Научно-техническая информация. Серия 2. М., 1970, № 8, 11-16.
Перебейнос В. И. (отв. ред.) Частотный словарь современной украинской художественной прозы. Пробная тетрадь. Киев, 1969.
Пиквер А. Частотный список морфем английского языка.- Linguistica, VII. Tartu, TRÜ, 1976, 177-189.
Ратцева И. И., Строганов В. А. О подстановках в тексте. - Вопросы структуры языка. Синтаксис, типология. М., "Наука", 1974, 119-132.
Ротарь А. С. Частотный словарь немецких публицистических текстов и его использование при обучении чтению. - Методические записки по вопросам преподавания иностранных языков в вузе. М., 1970, 135-158.
Частотный словарь сочетаемости современного английского языка. Под ред. Н. О. Волковой и др. М. 1972.
Шрейдер Ю. А. О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа). - Проблемы передачи информации. Том 3, вып. 1, 1967, 57-63.
Фрумкина Р. М. К вопросу о так называемом законе Ципфа. - "Вопросы языкознания", 1961, № 2.

T a b e l 1

Sõnavormi astak (r), esinemissagedus (F_r) ja selle järgsummad (F^{Σ}) ning suhteliste sageduste järgsummad (P^{Σ}) protsentides eesti ilukirjandusproosa autorikõne s õ n a v o r m i d e sagedussõnastiku põhjal (üld- ja pärisnimed koos)

r	F_r	F^{Σ}_r	P^{Σ}_r (%)	r	F_r	F^{Σ}_r	P^{Σ}_r (%)
1	3221	3221	3,22	200	48	35066	35,10
2	1602	4823	4,83	300	34	38592	38,63
3	1439	6262	6,27	400	25	41484	41,53
4	1375	7637	7,64	500	21	43755	43,80
5	1264	8901	8,91	600	17	45634	45,68
6	1116	10017	10,03	700	15	47264	47,31
7	995	11012	11,02	800	13	48703	48,75
8	713	11725	11,74	900	12	49966	50,02
9	592	12317	12,33	1000	11	51109	51,16
10	542	12859	12,87	1001–1034	11	51483	51,54
20	329	16801	16,82	1035–1168	10	52823	52,88
30	224	19344	19,36	1169–1299	9	54002	54,06
40	189	21384	21,41	1300–1500	8	55610	55,67
50	165	23130	23,15	1501–1753	7	57381	57,44
60	141	24635	24,66	1754–2131	6	59649	59,71
70	120	25920	25,95	2132–2650	5	62244	62,31
80	108	27045	27,07	2651–3460	4	65484	65,55
90	89	28017	28,05	3461–5088	3	70368	70,44
100	83	28871	28,90	5089–8973	2	78138	78,22
				8974–30733	1	99898	100,00

Tabel 2

Sõna astak (r), esinemissagédus (F_r) ja selle järgsummad (F^{*}) ning suhteliste sageduste järgsummad (P^{*}) protsentides eesti ilukirjandusproosa autorikõne lekseemide sagedussõnastiku põhjal (üld- ja pärisnimed koos)

r	F_r	F^{*}_r	P^{*}_r (%)	r	F_r	F^{*}_r	P^{*}_r (%)
1	4237	4237	4,24	200	71	50284	50,34
2	3493	7730	7,74	300	44	55785	55,84
3	2598	10328	10,34	400	33	59583	59,64
4	1981	12309	12,32	500	27	62517	62,58
5	1395	13704	13,72	600	22	64948	65,01
6	1300	15004	15,02	700	19	66986	67,05
7	1047	16051	16,07	800	16	68733	68,80
8	879	16930	16,95	900	14	70282	70,35
9	845	17775	17,79	1000	13	71622	71,70
10	827	18602	18,62	1001–1017	13	71843	71,92
20	448	24123	24,15	1018–1080	12	72599	72,67
30	309	27867	27,90	1081–1169	11	73578	73,65
40	268	30743	30,77	1170–1293	10	74818	74,89
50	223	33192	33,23	1294–1414	9	75907	75,98
60	190	35201	35,24	1415–1567	8	77131	77,21
70	174	36999	37,04	1568–1779	7	78615	78,70
80	153	38634	38,67	1780–2044	6	80205	80,29
90	135	40070	40,11	2045–2389	5	81930	82,01
100	124	41358	41,40	2390–2980	4	84294	84,38
				2981–3918	3	87108	87,20
				3919–5972	2	91216	91,31
				5973–14654	1	99898	100,00

Tabel 3

Seos sõnavormide sageduse (F_r) ja astaku (r) vahel; Zipfi valemi konstantide C, K ja γ arvutamine vähimruutude meetodi abil ilukirjandusproosa autorikõne koondsagedussõnastiku andmete põhjal

r	F_r	r	F_r	r	F_r
1	3221	20	329	1200	9
2	1602	30	224	1400	8
3	1439	40	189	1600	7
4	1375	50	165	1900	6
5	1264	100	83	2400	5
6	1116	200	47	3000	4
7	995	300	34	4200	3
8	713	500	21	7000	2
9	592	800	13	20000	1
10	542	1000	11	(n = 29)	-

Lineariseering:

$X = \lg r$ $\quad A = \lg C$ $\quad \sum X = 59,3540$

$Y = \lg F_r$ $\quad B = \gamma$ $\quad \sum Y = 54,0360$

$Y = A + BX$ $\quad \sum X^2 = 165,0093$

$\sum XY = 73,3379$

$$A = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n \sum X^2 - (\sum X)^2} = 3,6151,$$

$$B = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} = -0,8559.$$

$\lg F_r = \lg C + \gamma \lg r = 3,6151 - 0,8559 \lg r;$

$C = \text{antilog } A = 4122;$

$\gamma = -0,8559 \approx -0,86;$

$K = C/N = 4122/99898 = 0,0403.$

Seega: $F_r = 4122\ r^{-0,86}$

või

$P_r = 0,0403\ r^{-0,86}.$

T a b e l 4

Eesti tänapäeva ilukirjandusproosa autorikõne leksikaalne spekter
s õ n a v o r m i d e sagedussõnastiku andmete põhjal

F	Sõnastik				Tekst			
	m	m^{*}	p (%)	p^{*} (%)	mF	mF^{*}	p (%)	p^{*} (%)
1	21760	21760	70,8	70,8	21760	21760	21,8	21,8
2	3885	25645	12,6	83,4	7770	29530	7,8	29,6
3	1628	27273	5,3	88,7	4884	34414	4,9	34,5
4	810	28083	2,7	91,4	3240	37654	3,2	37,7
5	519	28602	1,7	93,1	2595	40249	2,6	40,3
6	378	28980	1,2	94,3	2268	42517	2,3	42,6
7	253	29233	0,8	95,1	1771	44288	1,8	44,4
8	201	29434	0,7	95,8	1608	45896	1,6	46,0
9	131	29565	0,4	96,2	1179	47075	1,2	47,2
10	134	29699	0,4	96,6	1340	48415	1,3	48,5
11 - 50	856	30555	2,8	99,4	17902	66317	17,9	66,4
51 - 100	96	30651	0,3	99,7	6637	72954	6,6	73,0
101 - 400	68	30719	0,2	99,9	12487	85441	12,5	85,5
> 400	14	30733	0,1	100,0	14457	99898	14,5	100,0
$\sum$	30733	-	100,0	-	99898	-	100,0	-

Tabel 5

Eesti tänapäeva ilukirjandusproosa autorikõne leksikaalne spekter
lekseemide sagedussõnastiku andmete põhjal

F	Sõnastik				Tekst			
	m	m^{*}	p (%)	p^{*} (%)	mF	mF^{*}	p (%)	p^{*} (%)
1	8682	8682	59,3	59,3	8682	8682	8,7	8,7
2	2054	10736	14,0	73,3	4108	12790	4,1	12,8
3	938	11674	6,4	79,7	2814	15604	2,8	15,6
4	591	12265	4,0	83,7	2364	17968	2,4	18,0
5	345	12610	2,4	86,1	1725	19693	1,7	19,7
6	265	12875	1,8	87,9	1590	21283	1,6	21,3
7	212	13087	1,4	89,3	1484	22767	1,5	22,8
8	153	13240	1,0	90,3	1224	23991	1,2	24,0
9	121	13361	0,8	91,1	1089	25080	1,1	25,1
10	124	13485	0,9	92,0	1240	26320	1,2	26,3
11 - 50	901	14386	6,1	98,1	19290	45610	19,3	45,6
51 - 100	144	14530	1,0	99,1	10303	55913	10,3	55,9
101 - 400	101	14631	0,7	99,8	18564	74477	18,6	74,5
> 400	23	14654	0,2	100,0	25421	99898	25,5	100,0
$\sum$	14654	-	100,0	-	99898	-	100,0	-

Tabel 6

Waringi-Herdani jaotuse kontrollimine hii-ruut-testi abil sõnavormide sagedussõnastiku materjali põhjal; F = sõnavormi sagedus, E - empiiriline ja T - teoreetiline esinemisarv; teksti maht N = 99898

F	E	T	E - T	$(E-T)^2/T$
1	(21760)	(21760)	-	-
2	3885	3872	+13	0,04
3	1628	1598	+30	0,56
4	810	861	-51	3,02
5	519	522	- 3	0,02
6	378	357	+21	1,24
7	253	258	- 5	0,10
8	201	194	+ 7	0,25
9	131	151	-20	2,65
10	134	120	+14	1,63
11	91	98	- 7	0,50
12	80	80	0	0
13	64	68	- 4	0,24
14	59	58	+ 1	0,02
15	50	49	+ 1	0,02
16	50	43	+ 7	1,14
17	50	40	+10	2,50
18	31	34	- 3	0,26
19	34	31	+ 3	0,29
20	20	28	- 8	2,29
21	21	24	- 3	0,38
22	32	22	+10	4,55
23	24	20	+ 4	0,80
24	21	18	+ 3	0,50
25	14	17	- 3	0,53
25	393	352	+41	4,78
n = 25	V = 30733	V^1 = 30733	-	χ^2 = 28,31

$$\chi^2 = 28,31 < \chi^2_{0,05;24} = 36,42$$

Tabel 7

Waringi-Herdani jaotuse kontrollimine hii-ruut-testi abil lekseemide sagedussõnastiku materjali põhjal; F - lekseemi sagedus, E - empiiriline ja T - teoreetiline esinemisarv; teksti maht N = 99898

F	E	T	E - T	$(E-T)^2/T$
1	(8682)	(8682)	-	-
2	2054	2308	-254	27,95
3	938	1049	-111	11,75
4	591	593	- 2	0,01
5	345	379	- 34	3,05
6	265	262	+ 3	0,03
7	212	191	+ 21	2,31
8	153	145	+ 8	0,44
9	121	113	+ 8	0,57
10	124	91	+ 33	11,97
11	89	73	+ 16	3,51
12	63	61	+ 2	0,07
13	77	50	+ 27	14,58
14	47	43	+ 4	0,37
15	37	37	0	0
n = 15				χ^2 = 76,61

$$\chi^2 = 76,61 > \chi^2_{0,05;14} = 23,68$$

ЧАСТОТНЫЙ СЛОВАРЬ КАК ОБ"ЕКТ ЛЕКСИКОСТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ

Ю. А. Тулдава

Р е з ю м е

В статье излагаются принципы составления частотных словарей (ЧС) и описывается процесс создания ЧС авторской речи современной эстонской художественной прозы. Основные параметры ЧС эстонского языка: объем текста N = 99898 словоупотреблений, объем словаря словоформ V = 30733, объем словаря лексем L = 14654. Полные данные о распределении частот слов приводятся в таблицах 1 и 2 (соответственно по словарю словоформ и словарю лексем; символы обозначают: r – ранг слова, F_r – частота его появления в тексте, F_r^+ – накопленная частота, P_r^+ – накопленная относительная частота в процентах). В статье подробно рассматриваются отдельные квантитативно-типологические параметры эстонского словаря и текста в сравнении с другими языками. Констатируется, что по схеме К. Бектаева (1975) эстонский язык попадает в группу флективно-синтетических языков (например, 100 самых частых словоформ покрывают 29 % эстонского текста при интервале покрываемости 24 ... 42 % для флективно-синтетических языков). Далее, рассматриваются некоторые важные лексикостатистические закономерности структуры словаря и текста. Описывается действие закона Ципфа и вычисляются константы Ципфа по данным ЧС эстонского языка: при N = 100.000 по словарю словоформ γ = 0,86 и K = 0,041 (с поправкой Мандельброта γ = 0,87, K = 0,046 и ς = 0,5). По словарю лексем: γ = 0,92 и K = 0,068. Констатируется хорошее соответствие эстонского текста теоретическому распределению по формуле Ципфа (см. рис. 3). Вычисляются и константы формулы Вейбулла-Белоногова: c = 0,053 и k = 0,379 (по словарю словоформ). Подробно рассматривается вопрос о лексическом спектре по данным ЧС (полные данные приводятся в таблицах 4 и 5 – соответственно по словарю словоформ и словарю лексем; символы обозначают: F – частота слова, m – число слов с данной частотой, m^+ – соответствующая накопленная частота, p^+ – накопленная относительная частота в процентах). Распределение частот лексического спектра проверяется по формуле Уэринга-Хердана и констатируется хорошее соответствие теоретическим данным, особенно на уровне словоформ (см. табл. 6).

THE FREQUENCY DICTIONARY AS AN OBJECT OF LEXICOSTATISTICAL INVESTIGATION

J. Tuldava

S u m m a r y

The article sets forth some basic principles of the compilation of frequency dictionaries (FD) and describes the process of the production of a FD of modern Estonian prose fiction (non-conversational material). The main parameters of this FD of Estonian: volume of the text N = 99,898 words ("tokens"), that of the vocabulary of word forms V = 30,733, that of the vocabulary of lexemes L = 14,654. Complete data on the frequency dictribution of words are presented in Tables 1 and 2 (on the vocabulary of word forms and that of lexemes, respectively; the symbols used denote: r - rank of the word, F_r - frequency of occurrence in the text, F_r^+ - accumulative frequency, P_r^+ - accumulative relative frequency in percentage). The article deals in detail with some quantitative-typological parameters of the Estonian vocabulary and text as compared with those of some other languages. It has been ascertained that according to K. Bektayev's scheme (1975) the Estonian language belongs to the group of inflectional-synthetical languages (e.g. 100 most frequent word forms cover 29 % of the Estonian text while the range of coverage for inflectional-synthetical languages is 24 ... 42 %). Some important lexicostatistical regularities of the structure of the vocabulary and text are likewise considered. The influence of Zipf's law is described and Zipf's constants calculated on the basis of data in the FD of Estonian: at N = 100,000 γ = 0.86 and K = 0.041 (taking into account Mandelbrot's correction γ = 0.87, K = 0.046 and ϱ = 0.5) for the vocabulary of word forms. For the vocabulary of the lexemes γ = 0.92 and K = 0.068. There is good agreement with the theoretical distribution according to Zipf's formula (see Fig. 3). The constants of Weibull-Belonogov's formula have also been calculated: c = 0.053 and k = 0.379 (for the vocabulary of word forms). The problem of the "lexical spectrum" (Yule's distribution) is discussed in detail on the basis of the FD (complete data are given in Tables 4 and 5 - for the vocabulary of word forms and that of lexemes, respectively; the symbols used denote: F - frequency of the word, m - number of the words with the given frequency, m^+ - corresponding accumulative frequency, p^+ - accumulative relative frequency in percentage). The distribution of frequencies of the lexical spectrum is checked by the Waring-Herdan formula and a good correspondence with theoretical data, especially on the level of word forms, is revealed (see Table 6).