EESTI KEELE SÕNAVARA FONEETILIS-GRAFEEMILISED MÕÕTED

Juhan Tuldava

Sõnavara statistilise vaatluse alla kuulub traditsiooniliselt ka sõnade välise kuju (koostis, pikkus) analüüs. Sõnu vaadeldakse sel juhul kui tähtede (grafeemide), häälikute või foneemide jadasid ja kombinatsioone, mille kvantitatiivne uurimine võimaldab kindlaks teha üksikute elementide produktiivsust ja funktsionaalset koormust ning kokkuvõttes luua sõnastiku ja teksti kvantitatiiv-struktuurne tüpoloogia foneetilis-grafeemilisel või fonoloogilisel tasandil. Jätkuks varem avaldatud töödele eesti keele sõnavara statistiliste omaduste kohta (vt. näit. Kaasik jt., 1976 ja 1977; Tuldava, 1977 ja 1978) vaatleme käesolevas artiklis eesti keele sõnavara foneetilis-grafeemilisi (häälikulisi ja tähelisi) mõõteid nii t e k s t i s kui s õ n a s t i k u s, kusjuures vaatlusmaterjal on võetud viiest eesti keele allkeelest (ilukirjandusproosa autorikõne ja tegelaskõne, ajaleht, teaduskirjandus, luule). Võrdluseks tuuakse näiteid teistest keeltest, esmajoones sugulas- ja naaberkeeltest.

Sissejuhatavad märkused. Foneetilis-grafeemiliste mõõdete all mõtleme laiemas mõttes nii häälikute või tähtede kui ka hääliku- või tähekombinatsioonide sagedusi teksti ja sõnastiku tasandil. Häälikute või foneemide kombinatsioonid ja nende põhjal saadud struktuurid kuuluvad kitsamas mõttes fonotaktika (tähtede puhul vastavalt grafemotaktika) valdkonda, mida antud juhul ei vaadelda (eesti keele kohta vt. Kaasik jt., 1975; Тулдава, 1978). Käesolevas töös käsitletakse vaid eesti keele tähtede ja häälikute, sealhulgas algus- ja lõpptähtede (-häälikute) sagedusi tekstis ja sõnastikus.

Nimetatud kvantitatiivsed vaatlused on vajalikud mitmesuguste rakenduslike ülesannete lahendamisel ja seepärast on vastavaid uurimusi tehtud väga paljude keelte kohta. Tähtede sagedusi on uuritud stenograafiliste süsteemide optimeerimiseks, kirjutusmasinate klaviatuuri koostamiseks, krüptograafias, logopeedias, tootenimede konstrueerimisel

jne. Olulist osa etendavad tähelised mõõted (sealhulgas tähekombinatsioonide sagedused) informatsiooni edastamisel sidekanalite kaudu (kusjuures tähtede ja nende kombinatsioonide sagedusi vaadeldakse informatsiooniteooria valguses), informatsiooni töötlemisel elektronarvuti abil jms. Tähelisti mõõdete vajalikkust näitab ka see, et nende abil on suudetud lahendada mõningad huvitavad lingvistilised probleemid. Tuntud on nõukogude teadlase B. Suhhotini katse klassifitseerida automaatselt tähti nende distributsiooni alusel (Сухотин, 1962); selle katse edasiarendamine võib viia ka foneemide automaatsele identifitseerimisele (Karlgren, 1968, 135). Võib nimetada ka soomlase S. Mustoneni (1965) katset määrata sõnade kuuluvust ühesse või teise keelde tähtede esinemissageduse alusel.

Tähelisi mõõteid on kasutatud ka keelte tüpoloogilisel võrdlemisel, kuigi sel juhul on õigem lähtuda häälikutest või foneemidest. Tuleb aga arvestada suuri raskusi keelte fonoloogilis-tüpoloogilisel uurimisel, mis on tingitud sellest, et sageli puudub ühtne teoreetiline alus eri keelte fonoloogiliseks kirjeldamiseks. Teatavasti puudub ühtsus isegi ühe keele fonoloogilise süsteemi (foneemivaru) esitamisel. Kujuka näite toob A. Isengeldina (Исенгельдина, 1972), kes on kindlaks teinud, et erinevates uurimustes kõigub inglise keele vokaalfoneemide arv 8 - 22 vahel. Ka eesti keele fonoloogilise süsteemi suhtes leidub vastandlikke vaateid (vt. Вийтсо, 1979), mistõttu käesolevas töös on loobutud eesti keele sõnade kvantitatiivsest uurimisest foneemide tasandil. Omaette alapeatükis vaadeldakse aga eesti keele häälikute ja nende rühmade sagedusi, lähtudes eesti keele traditsioonilisest häälikusüsteemist. See võimaldab teha tüpoloogilisi võrdlusi mõningate teiste keeltega. Kõrvutavad tüpoloogilised uurimused nii foneetilis-grafeemiliste kui ka fonoloogiliste mõõdete alusel etendavad tähtsat osa Leningradi teadlase N. Andrejevi keelestatistika koolkonna töödes (Андреев, 1967), kusjuures tähtede (ja foneemide) sagedusi sõnade eri positsioonides kasutatakse sõnade automaatsel morfoloogilisel segmenteerimisel. Huvitavate tulemusteni keelte kõrvutaval uurimisel foneetilis-grafeemiliste või fonoloogiliste mõõdete alusel on jõutud ka paljudes teistes töödes (näit. Krámský, 1959; Kučera, Francis, 1968;

Милецкий, 1963; Перебейнос, 1967; Вецовол, 1967).

Mitmed uurimused on käsitlenud tähe- ja häälikusageduste muutumist keelte ajaloolises arengus (Вектаев, Лобин, 1969; Зиндер, Строева, 1972). Eesti keele suhtes on seni kindlaks tehtud, et tähtede sagedused pole viimase sajandi jooksul oluliselt muutunud (Kaasik, Laugaste, 1969). Eesti sõnade kõige varasemad kirjapanekud lubavad aga järeldada, et vanema eesti keele häälikute süsteem oli küllaltki erinev tänapäevasest (Kask, 1972, 15). Ka vene keele suhtes on täheldatud olulisi muutusi eri ajajärkude häälikute ja foneemide süsteemides ning vastavates kvantitatiivsetes näitajates (Журавлев, 1974, 87). Huvi pakub rootsi teadlase B. Sigurdi (1963) uurimistulemus, mille kohaselt võib nentida foneemide üldarvu tõusu paljudes euroopa keeltes (näit. rootsi keeles on alates 13. sajandist foneemide arv tõusunud 25-lt 40-ni). B. Sigurd seletab seda kui "koodi nihet", nimelt liikumist sellise koodi poole, mis võimaldab edastada rohkem informatsiooni teatava ajaühiku vältel eeldusel, et foneemi keskmine vältus jääb samaks.

Paljusid uurijaid on huvitanud kardinaalne küsimus, miks üldse esinevad tähed (häälikud, foneemid) eri sagedusega. Informatsiooniteoreetilisest seisukohast võib seda seletada keele kui "koodi" optimaalse organisatsiooni tendentsiga, mis on täheldatav ka sõnasageduste erinevuse puhul (näit. Zipfi seadus). Miks aga esinevad just teatavad häälikud sagedamini kui teised, sellele pole suudetud anda lõplikku seletust. G. Zipf (1935) arvas, et sageduste erinevusi põhjustavad artikulatsiooni tingimused, nimelt olevat artikulatsiooni keerukus pöördvõrdeline hääliku sagedusega. Sellele on vastu vaieldud ja väidetud, et hääldusе "keerukust" või "kergust" on võimatu mõõta (Trubetzkoy, 1939; Martinet, 1955). V. Nikonov arvab, et hääliku või foneemi sageduse "võti" ei peitu üldse foneetikas või fonoloogias, vaid sõnavara ja sõnatuletuse ajaloolises arengus (Никонов, 1963). A. Isengeldina väidab siiski, toetudes konkreetsetele uurimustele, et peab olema mingi seos kõnefüsioloogia ja häälikusageduse vahel: nii näiteks võib täheldada peaaegu kõigis keeltes apikaalide suurt sagedust, mis seletub keeletipu liikuvuse ja aktiivsusega (Исенгельдина, 1972).

Kui toetuda analoogiale sõnade ja grammatiliste nähtuste sageduse ja "raskuse" seose kohta, siis peaks väitma, et antud keele kandjale on sagedamini esinev häälik "lihtsam" kui harva esinev häälik (nii käsitleb küsimust ka P. Guiraud, 1954, 97). Igal juhul märgib hääliku suur esinemissagedus ühtlasi suurt funktsionaalset koormust ja suuri kombinatoorseid võimalusi, mis kahtlemata on teatud määral seotud hääldusega lihtsusega (ökonoomiaprintsiip). Teisest küljest tuleb aga arvesse võtta häälikute tajumist ja eristamist kuulaja poolt, seda on viimasel ajal hakatud uurima ka psühholingvistiliste katsetega (näit. Меликишвили, 1970). Küsimuse lahendamisel võivad kasulikuks osutuda katsed subjektiivsete sageduste määramisel (Фрумкина и др., 1971). Nn. "foneetilise sümbolismi" pooldajad seostavad mõningatel juhtudel häälikute sagedusi antud keele kandja hinnanguga ("hea", "halb", "ilus" jne.) häälikule või häälikut sisaldavale sõnale (vt. Журавлев, 1974, 86 jj.). Peab nimetama, et foneetilise sümbolismi uurimine on viimastel aastatel saanud uut hoogu nii Nõukogude Liidus kui välismaal, kusjuures on kasutusele võetud uusi täiustatud (sealhulgas statistilisi ja psühholingvistilisi) meetodeid (Левицкий, 1969; Гурджиева, 1973; Журавлев, 1974; Marchand, 1959; Arndorfer, 1967; Ertel, 1969 jt.). Teatavasti uuris eesti keele häälikute omapära ja "tähendust", lähtudes vanakreeka filosoofide seisukohtadest, juba K.J. Peterson (vt. Ariste, 1968, 17 jj.). Uuemal ajal on tehtud katset seostada luuleteksti foneetilist struktuuri semantikaga (Mäger, 1971).

Tähtede sagedused tekstis. Paljude keelte kohta tehtud uurimused on näidanud, et tähtede esinemissagedus on küllaltki stabiilne antud keele või allkeele tekstides. Kõnelemisel või kirjutamisel oleneb tähtede või häälikute valik nendest sõnadest, mida me kõnes kasutame. Keeles kehtivad aga kindlad seaduspärasused sõnade häälikulise struktuuri suhtes ja seepärast ei saa me reeglina vabalt valida häälikuid, välja arvatud kunstlikult esilekutsutud juhtudel (kurioosumina võib nimetada inglase E. Wright'i katset kirjutada terve romaan ilma ühegi e-täheta või 18.sajandi saksa luuletaja G. Burmanni 130 poeemi ilma r-täheta). Tähtede sagedus on keelele omane nähtus, mis püsib konstantsena pi-

kema perioodi vältel, kuid täpsem analüüs näitab siiski mõningaid erinevusi allkeelte vahel. Sellest tingituna on uuemates uurimustes püstitatud nõue, et tähtede (samuti häälikute, foneemide) sagedusi vaadeldaks allkeelte kaupa. Juhul, kui soovitakse esitada andmed kogu keele kohta, tuleb ära näidata allkeelte doseering üldises valimis. Seejuures peetakse optimaalseks nelja-viie allkeele kaasatõmbamist "üldkeele" andmete kindlakstegemisel, näiteks järgmistes proportsioonides (Пиотровский, 1968, 44): ilukirjandusproosa - 30 %, kõnekeel (mida tavaliselt asendab tegelaskõne või draama) - 30 %, ajalehe- ja teaduskeel - 30 % ning luulekeel - 10 %. Umbes samasuguseid proportsioone kasutavad ka mõned välismaised autorid (näit. Kučera, Monroe, 1968: ilukirjanduskeel - 60 %, ajalehekeel - 20 %, teaduskeel - 10 %, luulekeel - 10 %).

Käesolev uurimus eesti keele tähtede sageduse kohta põhineb autori varasemal tööl (valimi maht N = 50.000 tähte; vt. Tuldava, 1970a), millele lisati ligi 30.000 tähte erinevates allkeeltest.+ Kokkuvõttes jaotuvad osavalimid järgmiselt:

ilukirjandusproosa	25420 tähte	(32 %)
kõnekeel (tegelaskõne)	20914 "	(27 %)
ajaleht	16208 "	(20 %)
teadusalane tekst	11358 "	(15 %)
luule	5000 "	(6 %)
kokku	78900 "	(100 %)

Tekstid pärinevad tänapäeva eesti keelest (ilmumisaeg pärast 1960.a.). Ilukirjandusproosa ühendab antud juhul nii autorikõne kui ka tegelaskõne (vahekorras umbes 80:20), kusjuures osavalimid on võetud 30 eri autori teostest. On uuritud ka autorikõnet ja tegelaskõnet omaette (vt. allpool). Kõnekeelt esindab antud katses ilukirjandusteoste tegelaskõne (peamiselt dialoog). Teadusalased tekstid on võetud enam-vähem võrdsetes proportsioonides keeleteaduslikest, ajaloolistest, loodusteaduslikest, matemaatilistest ja tehnika-alastest artiklitest (nii erialastest kui populaartea-

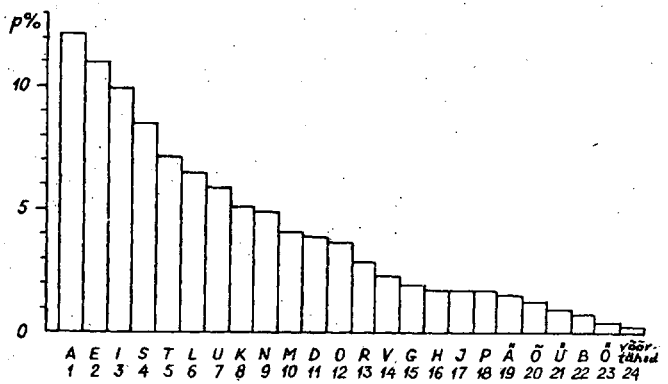
+ Käesolevas töös avaldatavad tähtede sagedused saadi valdavas osas tekstide automaattöötlemisel TRÜ arvutuskeskuses.

duslikest ajakirjadest). Luulekeele materjal (5·1000) pärineb erinevatest luulekogudest.

Koondvalimi proportsioonid on üldistavalt kokkuvõetuna järgmised:

ilukirjandus (proosa, tegelaskõne, luule)	65 %;
ajalehed ja teadusalased tekstid	35 %.

Uurimise tulemused esitatakse tabelis 1. Tabelis on toodud 23 eesti tähe ja võõrtähtede rühma (koos) esinemissagedused viie allkeele lõikes ja koondtekstis. Tähed järjestati sageduste kahanevas reas (koondteksti alusel). Joonis 1 illustreerib tähesageduste kahanevat rida histogrammi näol.



Joon. 1. Tähtede sagedused kahanevas reas (histogramm).

Koondteksti ("üldkeele") andmete järgi on kümme kõige sagedamat tähte ja vastavad suhtelised sagedused tekstis järgmised: a (12,2 %), e (10,9), i (9,8), s (8,6), t (7,1), l (6,5), u (5,8), k (5,1), n (4,9) ja m (4,1). Kõige harvemini esinevad tähed ü (0,9 %), b (0,8), š (0,3) ja võõrtähed (kokku ligikaudu 0,2 %). Vaadeldes võõrtähtede jaotumust eri allkeeltes (tabel 1)[+] näeme, et neid esineb kõige

[+] Hajuvushinnangud on antud tabelis 2 (arvutuste kohta vt. Tuldava, 1969).

rohkem ajalehekeeles (0,9 %), kuna aga kõigis teistes allkeeltes on võõrtähtede protsent 0 - 0,1 %. Huvi võib pakkuda võõrtähtede sageduste jaotumus meie koondvalimis (N = 78.900 tähte, sellest võõrtähti 203):

f	74	0,09 %	y	16	0,02 %
š	36	0,05	z	12	0,015
c	24	0,03	x	4	0,005
ž	18	0,02	q	2	0,0025
w	17	0,02			
			Kokku	203	0,25 %

Üllatab, et väike hulk sageli esinevaid tähti katab suure osa tekstist. Nii näiteks moodustavad 5 sagedamat tähte (a,e,i,s,t) koondtekstist 48,6 %, s.o. iga teine täht tekstis on üks nimetatud viiest. 10 sagedamat tähte katavad 75 % tekstist. Samasugune olukord valitseb ka teistes keeltes. Soome keeles katavad 10 sagedamat tähte isegi 80 % (Setälä, 1972, 11), hispaania keeles 76,2 %, inglise keeles 75,2 % (Pierce, 1962), saksa keeles 72,2 % (Meier, 1964, 334) tekstist.

Meie katse andmetel on eestikeelses tekstis v o k a a l t ä h t i keskmiselt 46,4 % ja k o n s o n a n t t ä h t i 53,6 %. Kui arvestada, et eesti tähestikus on 9 vokaalimärki ja 14 konsonandimärki, s.o. suhtes 39 % : 61 %, siis ilmneb, et vokaaltähtede "tekstikoormus" on tunduvalt suurem kui konsonantidel. Soomekeelses tekstis esineb vokaaltähti 49 % ja konsonanttähti 51 % (Setälä, 1972, 11), ungari keeles vastavalt 39 % ja 61 % (Статистико-комбинаторное моделирование, 1965, 206).

Sageduste hajuvuse analüüs usalduspiiride järgi näitab, et eestikeelsest tekstis võib kindlasti lugeda kõige sagedamateks tähti a, e, i, s, kusjuures viimane neist paikneb alati neljandal kohal. Viiendal kohal võib olla t või l. Hajuvust eri allkeelte vahel väljendab kõige paremini suhteline näitaja - suhteline viga (vt. tabel 2).

Erinevused tähtede sagedustes peegeldavad erinevusi sõnavaras, eriti sageli esinevates sõnavormides. Nii näiteks esinevad sõnad (sõnavormid) ja, ta, ka ilukirjandusproosas ja kõnekeeles eriti sageli ja seetõttu suureneb nendes allkeeltes ka a-tähe osatähtsus. Kõnekeeles (tegelaskõnes) ka-

sutatakse palju 1. isiku asesõna vorme (mina, ma, minu, meie, meid jne.), mis loomulikult suurendab m-tähe sagedust. Tabelist 1 on näha, et m-tähte esineb kõnekeeles 4,9 %, s.t. oluliselt üle keskmise. Selgub samuti, et t-tähte esineb palju ajaleht- ja teaduskeeles, kuna aga luuletekstides on t-tähe sagedus oluliselt alla keskmise. Seevastu esinevad luules tähed n ja l sagedamini kui teistes allkeeltes. Arvutused näitavad, et a-täht on kõige sagedam neljas allkeeles viiest, kuid teadusalastes tekstides paikneb esikohal e. Seda kinnitab ka H. Holmi (Хольм, 1965a) uurimus tähtede sagoduste kohta eesti tehnika-alastes tekstides. Meie katsematerjali lähemal analüüsimisel selgub, et teadusalastes tekstides on e-täht keskmiselt kaheksal juhul kümnest sageduselt esikohal. See seletub teaduskeele eripäraga, nimelt esineb selle allkeele tekstides palju omastavat käänet lõpphäälikuga e- (näit.: ... uue nõukogude kirjanduse tekkimine ..., tugevate mõjude), alaleütlevat käänet lõpuga -le (tavaliselt olukorrale, keskselt kohale), sõnu ja sõnade ühendeid nagu selline, selleks et, sellega seoses, seetõttu, sel teel jne. Teadusalasele tekstile on iseloomulikud ka adjektiivid liitega -ne (konkreetne, negatiivne, deskriptiivne), substantiivid liitega -mine (arenemine, tootmine), rahvusvahelised verbid liitega -seeri- (defineerima, demonstreerima, reprodutseerima) ja muud laen- ning võõrsõnad (probleem, protsess, meetod jt.).

Allkeeltevahelisi erinevusi saab kindlaks teha ka tähestike otsese kõrvutamisega. Eespool vaatlesime ilukirjandusproosat terviklikult, s.o. autori- ja tegelaskõnet koos. Üksikute tähtede sagedused erinevad aga oluliselt a u t o r i- ja t e g e l a s k õ n e s. Seda tõendab meie eriuurimus (vt. tabel 3), millest selgub, et näit. tegelaskõnes võib sagedamini kohata tähti e, m ja n, autorikõnes aga tähti u ja p (vahe olulisust on kontrollitud z-testi abil; vt. Tuldava, 1970b, 137-138). Täheliste mõõdete erinevus ilmneb ka selles, et vokaaltähti on autorikõnes 45,4, tegelaskõnes aga 47,4 %. Samuti võib täheldada olulist erinevust sagedate tähtede kontsentratsioonis: 5 sagedamat tähte katavad autorikõne tekstist 47,8 %, tegelaskõne tekstist aga 49,6 %.

Allkeeli võib võrrelda ka teisest seisukohast - läh-

tudes vastavate tähestikkude omavahelisest korrelatsioonist. Korrelatsioonikordaja (r) rakendamine toob esile suure lähheduse kõigi allkeelte tähestike kui tervikute vahel:

	Allkeel	II	III	IV	V	VI
I	Autorikõne	0,986	0,986	0,968	0,987	0,994
II	Tegelaskõne	-	0,984	0,976	0,979	0,996
III	Ajaleht		-	0,985	0,975	0,994
IV	Teadus			-	0,950	0,984
V	Luule				-	0,985
VI	Üldkeel					-

Antud korrelatsioonimaatriksit on võimalik kasutada allkeelte rühmitamiseks dots. L. Võhandu meetodi järgi (vt: Lepamaa, 1968, 101 jj.). Tulemused esitame skeemi kujul:

Luule - Autorikõne - Üldkeel - Ajaleht - Teadustekst
Tegelaskõne

Näeme, et moodustub kolm rühma, mis grupeeruvad üldkeele+ ümber: 1) ilukirjandusproosa autorikõne ja luule, 2) tegelaskõne, 3) ajaleht ja teadus- ning tehnika-alased tekstid. Kvantitatiivse meetodi abil saadud hinnangud vastavad intuitiivsele kujutlusele ja kvalitatiivse analüüsi tulemusena kujunenud rühmitustele; näiteks eristab E. Tissenko (Тисенко, 1971, 353) järgmisi grupeeringuid: 1) ajaleht ja teadustekstid, 2) dramaturgia, 3) ilukirjandus. Käesoleva katse andmetel moodustavad tuumiku ilukirjandusproosa autori- ja tegelaskõne ning ajalehekeel, mis on kõige lähemal koondtekstile ja ka omavahel lähedalt seotud. Perifeerias ja teineteisest kaugel asuvad luulekeel ja teaduskeel. Esimene neist on kõige tihedamini seotud ilukirjandusproosa autorikõnega, teine seostub kogusüsteemiga ajalehekeele vahendusel.

Käesoleva eksperimendi tulemused näitavad, et kuigi eri allkeelte tähelised mõõted on antud keele piirides omavahelises korrelatsioonis, mis lubab kõnelda teatud keelele kui tervikule omasest "graafilis-tähestikulisest struktuu-

+ Üldkeele all mõeldakse antud juhul allkeelte summat, mis puhul saadakse keele keskmised kvantitatiivsed näitajad (koondteksti andmed).

rist" (Тисенко, 1971), võib siiski vastavate statistiliste meetodite abil kindlaks teha süsteemisiseseid erinevusi allkeelte vahel. Küllalt suurte valimite korral peaks olema võimalik teksti tähtsageduste alusel allkeeli a u t o m a a t s e l t i d e n t i f i t s e e r i d a. Allkeelte ja stiilide vahelisi erinevusi on suudetud tähtede sageduste abil kindlaks määrata ka mõningates teistes keelestatistilistes uurimustes (näit. ukraina keele suhtes on olulised erinevused allkeelte vahel kindlaks tehtud foneemide tasandil, vt. Статистичні параметри, 1967, 44 jj.). Katsete tulemused kummutavad seega laialt levinud seisukoha, mille järgi tähtede sagedused on antud keele kõigi allkeelte ulatuses täiesti stabiilsed (statistiliselt homogeensed).

Lõpuks mõni sõna varasematest uurimustest tähtede sageduste kohta eestikeelses tekstis (uurimusi häälikute sageduste kohta vaatleme hiljem). Esimene teadaolev uurimus pärineb H. Hansenilt (1961), kes vaatles eestikeelses tekstis esinevate väikeste trükitähtede sagedusi trükikoja ladumismasinamatriitside ratsionaalsema komplekteerimise huvides. Järgnevad H. Holmi (Хольм, 1965a, 1965б) uurimused tähtede sageduste kohta eestikeelsetes raadioelektroonika- ja ajalehetekstides. 1969.a. avaldati "Keeles ja Kirjanduses" Ü. Kaasiku ja E. Laugaste elektronarvuti abiga teostatud uurimuse tulemused tähtede esinemissageduse kohta vanemas ja uuemas ajalehekeeles, ilukirjanduses ja rahvalaulus. Seda uurimust täiendati uute andmetega 1975.a. (Kaasik, Laugaste, Ääremaa, 1975). Käesoleva töö autor avaldas 1970.a. andmed viie allkeele tähtede sageduste kohta artiklis "Informatsiooniteooria ja keeleteadus" (Tuldava, 1970a). Kõigi nimetatud uurimuste tulemused, kaasa arvatud ka käesoleva töö andmed, on üksteisele küllaltki lähedased, ja erinevusi võib seletada sellega, et valimite koostis (allkeelte proportsioonid) on olnud eri uurimuste puhul erinev. Tabelis 4 esitatakse H. Hanseni, Ü. Kaasiku, E. Laugaste ja K. Ääremaa uurimuste tulemused koos käesoleva katse andmetega. Tabeli viimases veerus võetakse kokku Ü. Kaasiku, E. Laugaste ja K. Ääremaa andmed (tänapäeva eesti keele kohta) ja käesoleva töö autori uurimuse tulemused. See võimaldab esitada eestikeelse teksti tähtede

sagedused suurema valimi (üldmaht N = 163.340 tähte) põhjal ja järgmistes proportsioonides:

I:	ilukirjandusproosa	65970 tähte	- 40 %
	kõnekeel (tegelaskõne)	20914 "	- 13 %
	luule	5000 "	- 3 %
	ilukirjandus kokku	91884 "	- 56 %
II:	ajaleht	60098 "	- 37 %
	teadus- ja tehnikakeel	11358 "	- 7 %
	"tarbekirjandus" kokku	71456 "	- 44 %
	K o o n d t e k s t	163340 tähte	- 100 %

Sellises "ühendatud" valimis on 10 sagedamat tähte järgmised: a (12,6 %), e (11,0), i (9,6), s (8,7), t (7,1), l (6,3), u (5,9) k (5,0), n (4,8), d (4,0). Järgneb m (3,9), mis meie koondvalimi andmetel (vt. 5 veerg tabelis 5) on 10. kohal enne d-tähte. Ühendatud valimis katavad 5 sagedamat tähte 49,0 % ja 10 sagedamat tähte 75,0 % tekstist; vokaaltähti on 46,6 %, konsonanttähti 53,4 %. Need andmed on lähedased meie koondvalimi vastavatele näitajatele (48,6 - 75,0 - 46,4 - 53,6; vt. eespool).

Ü. Kaasiku, E. Laugaste ja K. Ääremaa uurimustes (1971 ja 1975) on esitatud huvitavad andmed tähtede sageduse kohta vanemas eesti kirjakeeles ja rahvalaulus. Vanemat kirjakeelt esindavad 5 kirjutist 1860-ndate aastate "Eesti Postimehest" vanas kirjaviisis (valimi maht N = 47700 tähte); rahvalaulu tekstid on võetud J. Hurda "Vana kandle" II köitest ja antoloogiast "Eesti rahvalaulud" (Tallinn, 1969) üldmahuga N = 43250 tähte. Ka vanemas ajalehekeeles ja rahvalauludes on 10 sagedamat tähte samad mis tänapäevakeeles (10. ja 11. kohal vahelduvad m ja d), kusjuures 10 tähe tekstikatmus on vanemas ajalehekeeles 74,5 ja rahvalauludes 76,2 %. Vokaaltähti on vastavalt 47,4 ja 48,3 %. Üldjoontes on tähtede sagedused vanemas kirjakeeles ja rahvalaulus küllaltki lähedased tänapäevakeelele, kusjuures vanem ajalehekeel läheneb enam tänapäeva ilukirjandus- ja kõnekeelele kui tänapäeva ajalehekeelele. Nii vanemas ajalehekeeles kui ka rahvalauludes torkab silma a- ja ä-tähe suur sagedus võrreldes tänapäevakeelega, kuna aga o-tähte on oluliselt vähem; rahvalauludes esineb suhteliselt palju l-tähte, suh-

teliselt vähe aga t-tähte (selle poolest on rahvalaul lähedane tänapäeva luulele). Öeldut illustreerib väljavõte tabelist 4:

	Vanem ajalehekeel	Rahvalaul	Tänapäevakeel
a	13,8 %	13,7 %	12,6 %
ä	2,3 %	2,7 %	1,4 %
o	2,8 %	2,8 %	3,6 %
t	6,9 %	5,7 %	7,1 %
l	5,6 %	7,1 %	6,3 %

.x.x.x.x.

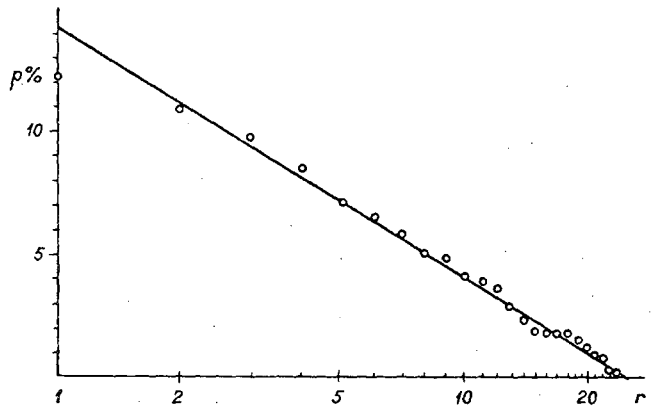
Eespool nägime, et tähtede sagedused moodustavad kõigis allkeeltes enam-vähem ühtlaselt langeva rea. See nähtus on omane üldse kõigile teadaolevatele keeltele. Uurijad on tundnud huvi sobiva funktsiooni vastu, mis väljendaks seost tähesageduse ja astaku vahel (samuti nagu Zipfi seaduse puhul sõnade sageduse ja astaku vahekorra uurimisel). Prantslased R. Moreau (1963) ja A. Verglas (1963) näitasid esimestena, et seos tähesageduse ja astaku vahel on antud keele puhul konstantse iseloomuga. Ameeriklane A. Mackay (1965) formuleeris seaduspärasuse logaritmilise regressioonvõrrandi näol: $p_r = a + b \lg r$, s.t. tähe suhteline sagedus (p_r) on lineaarses seoses astaku logaritmiga (lg r), kusjuures a ja b on konstandid. Kontrollides funktsiooni eesti keele andmete põhjal (võttes aluseks tänapäeva keele "suure" koondteksti, vt. 6. veerg tabelis 4), saame tulemuse

$$p_r = 13,7 - 9,6 \lg r.$$

Joonisel 2 on kujutatud logaritmilise horisontaalskaalaga diagramm, mis näitab küllaltki head lineaarset seost tähtede sageduse ja astaku logaritmi vahel, kuigi peab nentima, et kõige sagedam täht hälbib mõnevõrra üldisest tendentsist. Soome keele kohta (Setälä, 1972, 11) on saadud sellele lähedane tulemus:

$p_r = 14{,}4 - 10{,}3 \lg r$. Vene keele andmeil (Курбатов, 1968, 161) on vastav funktsioon:

$p_r = 11{,}4 - 7{,}5 \lg r$.*



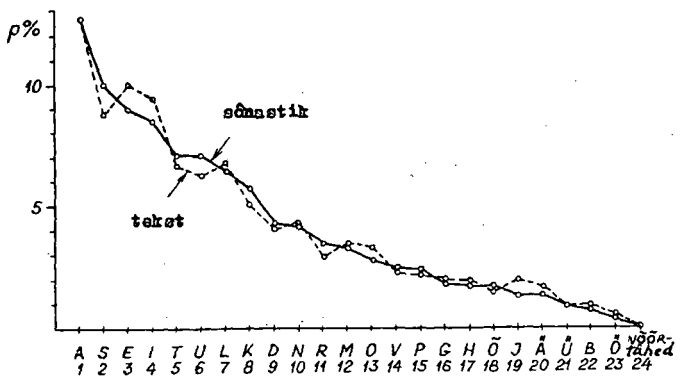
Joon. 2. Seos tähtede sageduse ja astaku vahel (poollogaritmdiagramm).

Tähtede sagedused sõnastikus. Tähelisi mõõteid võib uurida nii tekstis kui ka sõnastikus. Sõnastikusageduste uurimisel kerkib aga alati küsimus, missugune sõnastik võtta aluseks foneetiliste üksuste vaatlemisel. Kui lähtuda tavalisest lekseemide sõnaraamatust, milles on antud ainult sõnade põhivormid (verbi infinitiiv, substantiivi nimetav käänne jne.), siis peab arvestama, et põhivorm on suvaliselt (kokkuleppeliselt) valitud sõnavorm. Foneetiliste üksuste sageduspilt esineb lekseemide sõnastikus mõnel määral moonutatud kujul, võrreldes tavalise kõnega või tekstiga, näit.

* Kui võtta arvesse ainult 15–20 sagedamat tähte, siis saab tähtede sageduse ja astaku vahelist seost kõigi nimetatud keelte puhul veelgi täpsemalt väljendada eksponentfunktsiooni $p_r = Ae^{-Br}$ abil, kus A ja B on konstandid. Eesti keele suhtes (20 sagedama tähe põhjal) kehtib $p_r = 14{,}2\ e^{-0{,}13r}$, soome keele puhul $p_r = 15{,}7\ e^{-0{,}13r}$ ja vene keele puhul $p_r = 11{,}2\ e^{-0{,}1r}$.

infinitiivvormide suure esinemissageduse tõttu tõuseb ülemäära tähtede m ja s sagedus (tunnuse -ma arvel), nimetava käände ühekülgne domineerimine toob endaga kaasa konsonantide sageduse suurenemise (lõpptähe arvel) jne. Järelikult ei ole päris õige lähtuda täheliste või foneetiliste (fonoloogiliste) üksuste statistilisel uurimisel lekseemide sõnastikust (veel enam kehtib see väide sõnapikkuste ja fonotaktiliste nähtuste uurimise puhul). Objektiivsema pildi saame, kui vaatleme tähelisi mõõteid s õ n a v o r m i d e s õ n a s t i k u s. Kuid igal juhul tuleb arvesse võtta sõnastiku mahtu. Peab silmas pidama, et suurtes sõnastikkudes esineb tavalisest rohkem perifeerset sõnavara, näit. eriala- ja võõrsõnu, mille foneetiline koostis erineb igapäevaste sõnade omast. Uurimisel peab seepärast eelnevalt kokku leppima sõnastiku liigi ja mahu suhtes, eriti kui on tegemist võrdleva tüpoloogilise käsitlusega.

Käesolevas töös vaadeldakse paralleelselt kitsapiirilise allkeele teksti ja sõnastikku. On valitud tänapäeva eesti i l u k i r j a n d u s p r o o s a a u t o r i k õ n e, millest on tehtud tekstivalim 10420 tähe ulatuses ja sõnastikuvalim sõnavormide tasandil üldmahuga 19770 tähte (valimid on võetud 10 eri autori teostest). Tabelis 5 esitame tähtede sagedused tekstis ja sõnastikus ja sageduste vahed. Joonisel 3 on kujutatud graafiliselt teksti- ja sõnastikusagedused viimaste kahaneva rea järgi.



Joon. 3. Tähtede sagedused sõnastikus ja tekstis.

Autorikõne sõnavormide sõnastikus on sagedamad tähed järgmised: a (12,8 %), s (10,0), e (9,0), i (8,5), t (7,1), u (7,1), l (6,5), k (5,7), d (4,3), n (4,2). Olulisim erinevus, võrreldes tekstisagedustega, on nähtavasti s-tähe nihkumine teisele kohale sagedusjärjestuses; sageduste vahe on 1,2 % sõnastiku kasuks (vt. tabel 5), mis on statistiliselt oluline. Sisuliselt tähendab see seda, et s-täht esineb "laialipaisatult" väga paljudes erinevates sõnades (sõnavormides) ja ei ole kontsentreerunud teatava väiksema rühma sõnadesse. Sama võib öelda u, k, r, õ kohta, kuigi vähemal määral (teksti- ja sõnastikusageduse vahe on vähemalt 0,3 %). Nende tähtede tekstikoormus (funktsionaalne koormus teksti tasandil) on suhteliselt väike, kuid "informatiivsus" suur, s.t. nende esinemine tekstis signaliseerib keskmiselt rohkem erinevate sõnade või vormide ilmumisest kui teised tähed. Suhteliselt suure tekstikoormusega on aga tähed e, i, l, o, j, ä, mille tekstisagedused ületavad sõnastikusagedusi. Kui mõõta tekstikoormust tekstisageduse (T) ja sõnastikusageduse (S) suhtega, siis näeme, et suurim T/S väärtus on j-tähel (2,0/1,4 = 1,43). See seletub peamiselt sõna ja suure sagedusega tekstis (sõnastikus üksainus vorm). Ka e ja i tekstikoormust mõjustab suurel määral sõnavormi ei suur sagedus. Tähe ä suhtes tuleb nentida, et sõnavarasse kuulub võrdlemisi vähe ä-lisi sõnu, kuid mõni neist esineb tekstis sageli, näit. ära, pärast, välja, jälle, läks.

Vaadeldes vokaalide ja konsonantide sageduste suhet, näeme, et tekstis on suhe 46,4:53,6 ja sõnastikus 44,6:55,4. Järelikult on vokaalide osatähtsus tekstis mõnel määral suurem kui sõnastikus.

Viis sagedamat tähte katavad autorikõne tekstist 47,8 % ja sõnastikust 47,3 %, kümme sagedamat tähte aga vastavalt 74,4 % ja 75,1 %. Need arvud on küllaltki lähedased. Ka tähestikusageduste üldine struktuur (vt. joon. 3) on samalaadne. Teksti- ja sõnastikusageduste vaheline korrelatsioon on erakordselt kõrge (r = 0,988). Võib teha üldise järelduse, et teksti ja sellele tekstile vastava sõnastiku sagedused on tõepoolest väga lähedased. Üksikute tähtede puhul ilmnevad mõningad erinevused, mis seletuvad sellega, et näit. rida sõnu esineb tekstis suure sagedusega, kuna ago sõnastikus on neil ainult üks koht. Selle ebaproportsionaal-

suse arvel toimub ka mõningate tähtede sageduse suurenemine tekstis, võrreldes vastava sõnastikuga (ja ümberpöördult).

Häälikute sagedused tekstis. Eesti keeles on traditsioonilise arvestuse järgi 20 häälikut: 9 täishäälikut ja 11 kaashäälikut (tugev ja nõrk sulghäälik loetakse üheks häälikuks, seega k+G, p+B, t+D annavad kolm häälikut; kokku arvestatakse ka n+ŋ).+ Peale selle võivad esineda võõrhäälikud (f, š ja helilised ž, ǯ, g, b, d), murdeti ka rida teisi häälikuid (vt. Ariste, 1968, 90-91). Ühendeid ts, tš, dz, dž käsitletakse vahel afrikaatidena, kuid silmas pidades seika, et eesti keeles paikneb silbipiir keset afrikaati (täpsemalt keset afrikaadi sulghäälikulist osa) võime nimetatud ühendites arvestada kahte häälikut.

Häälikute sagedusi eestikeelses tekstis uuris esimesena A. Saareste (1932), kuid kahjuks on tema andmed trükis avaldatud ainult osaliselt ja pole ka teada, kui suure valimi ja mis laadi tekstide alusel sooritati häälikute loendus. Põhjalikuma uurimuse teostas arstiteadlane V. Särgava (1967), kes vaatles häälikute sagedusi tänapäeva ajaleheja ajakirjatekstides üldmahuga umbes 30.000 häälikut (uurimuse tulemusi kasutati kõneaudiomeetriliste katsete läbiviimisel).

Käesolev töö on mõeldud varasemate uurimuste täienduseks, kusjuures valimi materjali on mitmekesistatud ja analüüsitud lingvistilisest seisukohast. Valim on koostatud 4 allkeele põhjal (ilukirjandusproosa autorikõne, tegelaskõne, ajaleheja populaarteaduslik tekst - kõik võrdsetes osades), valimi üldmaht, s.o. teksti pikkus N = 23.560 häälikut. Tabelis 6 on toodud nii meie katse tulemused kui ka A. Saareste ja V. Särgava andmed häälikute sageduse kohta eestikeelses tekstis ("kirjalikus kõnes"). Kõigi kolme katse tulemused langevad üldjoontes kokku. Väikesed erinevused üksikute häälikute sagedustes (näit. on käesoleva töö autori läbiviidud katse andmeil häälikut t/D rohkem ja k/G vähem kui V. Särgava uurimuse tulemuste järgi) tulenevad ilmselt valimite erinevast koostisest allkeelte lõikes. Ka

---

+ Leenisklusiilid G, B ja D esinevad kirjas kui g, b, d. Kombinatoorset nasaali ŋ eesti keele kirjas eri märgiga ei tähistata.

häälikute puhul peab arvestama mõningaid erinevusi allkeelte vahel, nagu seda võisime nentida tähtede sageduste uurimisel.

Eesti keele h ä ä l i k u d ja t ä h e d on teatavasti tihedas korrelatsioonis, kusjuures peamine erinevus seisneb selles, et pikki häälikuid väljendatakse vastavate tähtede kahekordse kirjutamise teel. Sellest tulenevalt on ühe ja sama teksti ulatuses alati rohkem tähti kui häälikuid. Antud katse puhul oli valimis 24 796 tähte ja 23 560 häälikut, seega on suhe 105:100. Võrdluseks toome andmed mõnede teiste keelte kohta (Meier, 1964, 321): itaalia keel - 104:100, saksa keel 112:100, hollandi keel - 114:100, taani keel - 124:100, prantsuse keel - 148:100. Nagu näha, on eesti keel lähedane itaalia keelele ortograafia ökonoomsuse poolest. Pikkade häälikute kahekordne kirjutamine eesti keeles ei koorma seega kuigivõrd meie õigekirjutussüsteemi.

Tabelis 7 esitame uuesti meie katse andmed häälikute sageduse kohta, märkides eraldi ühe ja kahe tähega kirjutatud häälikute sagedused. Kokkuvõttes on eestikeelses tekstis lühikesi, s.o. ühe tähega kirjutatud vokaale 92 % ja pikki, s.o. kahe tähega kirjutatud vokaale - 8 %. Soome keeles on lühikeste ja pikkade vokaalide suhe samuti 92:8 (arvestused V. Setälä uurimuse andmete põhjal). Ühe tähega kirjutatud konsonanthäälikuid on eestikeelses tekstis 97 % ja kahe tähega kirjutatud konsonante - 3 %. Soome keeles on vastav suhe - 90:10. Siinjuures tuleb aga arvestada, et eesti keeles on tegelikult pikad ka II- ja III-vältelised geminaatklusiilid, mis kirjas märgitakse ühe tähega (näit. lepib [leppiB] , seepi [sēppi]), samuti sõnalõpulised ühe tähega kirjutatud fortisklusiilid (näit. kompvek, taburet). Nimetatud juhtude kaasahaaramisega suureneks "pikkade" konsonanthäälikute osakaal ligikaudu 13 %-ni (kõigist konsonanthäälikutest tekstis).

K õ i g e s a g e d a m a d h ä ä l i k u d eestikeelses tekstis on käesoleva uurimuse andmeil a (12,2 %), t/D (11,9), e (11,0), i (9,5), s (9,0). Järgnevad k/G, l, u, n/ŋ, m, o, r, p/B, v. Väikese sagedusega (igaüks alla 2%) on j, h, ä, õ, ü, ö ja võõrhäälikud. Võrdluseks toome viis sagedamat häälikut soome keeles (Setälä, 1972, 38):

a (11,6 %), n (10,2), i (10,2), e (9,4), t (8,9). Viis sagedamat häälikut katavad eestikeelsest tekstist 53,6 %, soomekeelsest 50,3 %. Seega mõjustavad sagedamad häälikud küllaltki oluliselt hääldust üldpilti. Võrreldes mõlemat sugulaskeelt, näeme, et kuigi viie sagedama hääliku hulgas on neli ühist (a, t, e, i), seisneb erinevus selles, et soomekeelses kõnes on n-häälik suure esinemissagedusega (eesti keeles on n 9. kohal sagedusega 4,3 % soome 10,2 % vastu), kuna aga s- ja t-häälikut esineb soome keeles oluliselt harvemini (s-hääliku sagedus on soome keeles 7,0 %).

Viis sagedamat häälikut ungari keeles on L. Hakulinen-i (1941) andmeil: a (14 %), a (13 %), t (8 %), l (6 %), n (5,5 %). Mansi keeles reastab B. Kálmán (1963) viis sagedamat häälikut järgmiselt: a (10,1 %), t (9,0 %), i (7,0 %), l (5,8 %), m (5,3 %). Kõigile ülalnimetatud soome-ugri keeltele on ühised suure sagedusega häälikud a ja t (viie sagedama hääliku hulgas).

Huvipakkuv on vaadelda v o k a a l i d e ja k o n s o n a n t i d e suhet eesti keeles, võrreldes teiste keeltega. Meie koondvalimi (4 allkeelt) andmeil esineb eestikeelses tekstis vokaalhäälikuid 45,5 % ja konsonanthäälikuid 54,5 %, seega 100 vokaali kohta 120 konsonanti (A. Saareste andmeil 117). Soome keeles on täis- ja kaashäälikute suhe 48,2:51,8, s.o. 100 vokaali kohta 108 konsonanti.

Täis- ja kaashäälikute sageduste suhet on alati peetud tähtsaks karakteristikuks keelte tüpoloogilisel uurimisel. Juba B. Bourdon (1892) jaotas keeled vokaalseteks, konsonantseteks ja "segatüüpi" keelteks. J. Krámský (1948) mõõtis keelte "vokaalsust" erilise koefitsiendi abil: v = $= p_i/p_t$, kus p_i tähistab kaashäälikute protsenti süsteemis (häälikuvarus) ja p_t - kaashäälikute protsenti tekstis. Kui arvestada ainult eesti oma häälikuid (11 konsonanti 20 hääliku kohta), saame p_i = 55 % ja vokaalsuskoefitsiendi v = 55:54,5 = 1,01. Soome keeles on 13 konsonanti 21 hääliku kohta, seega p_i = 62 % ja v = 62:51,8 = 1,20. Võrdluseks võib tuua J. Krámský andmed mõne teise keele kohta: saksa keel - 0,85, inglise keel - 0,91, hispaania keel - 1,46, itaalia keel - 1,58. Antud juhul on nähtavasti õigem arvestada eesti keeles ka suhteliselt kodunenud võõrhääli-

kuid (näit. f ja š). Konsonante oleks siis kokku 13 ehk 59 % häälikute üldarvust (22). Vokaalsuskoefitsiendi väärtuseks saame v = 59:54,5 = 1,07. Veelgi kõrgema vokaalsuse hinnangu saaksime, kui arvestaksime kõiki eesti keeles esinevaid võõrhäälikuid.

Keelte "vokaalsusastet" võib objektiivselt määrata ka tekstisageduste põhjal. Võttes aluseks suhte p_v/p_k, kus p_v tähistab vokaalide ja p_k konsonantide sagedust tekstis, arvutame eri keelte vokaalsusastme $v_t = p_v/p_k$ (sulgudes märgime konsonantide arvu 100 vokaali kohta tekstis), ja jaotame keeled vokaalsusastme järgi kolme rühma, näiteks:

$v_t > 1$: tahiiti 2,40 (41), havai 1,54 (65), jaapani 1,22 (82);

$0,70 \leq v_t \leq 1$: portugali 0,98 (102), ruumeenia 0,94 (106), itaalia 0,93 (108), soome 0,93 (108), esperanto 0,87 (115), prantsuse 0,87 (116), hispaania 0,85 (118), eesti 0,83 (120), leedu 0,78 (129), läti 0,72 (138), ungari 0,71 (141);

$v_t < 0,70$: mansi 0,67 (149), poola 0,67 (148), vene 0,67 (150), slovaki 0,67 (150), inglise 0,63 (158); rootsi 0,63 (158), saksa 0,61 (164).

Andmed häälikute sageduste kohta on võetud eri uurimustest (Weiss, 1961; Meier, 1964; Zsilka, 1971; Setälä, 1972; Сапожников, 1966; Венсов, 1973). Antud juhul on tegemist peamiselt ilukirjandusproosa või ajalehetekstidega. Ei tohi aga unustada, et vokaalide ja konsonantide sagedused võivad kõikuda allkeelte kaupa. Nii näiteks esineb M. Weissi (1961) andmeil rootsi ajalehetekstis 160 konsonanti 100 vokaali kohta (vokaalsusaste $v_t = 0,63$), kuna aga telefonikõnedes on vastav sagedussuhe 138:100 ($v_t = 0,72$). Vokaalide osatähtsus tõuseb seega rootsi kõnekeeles märgatavalt, võrreldes kirjutatud tekstiga. Sama kehtib ka eesti keele kohta. Meie uurimistöö andmeil esineb tegelaskõne osavalimis täishäälikuid 46,8 % ja kaashäälikuid 53,2% (koondvalimis oli vahekord 45,5:54,5). See tähendab, et 100 vokaali kohta tuleb kõnekeelt imiteerivas ilukirjandusproosa tegelaskõnes 114 konsonanti (vokaalsusaste $v_t = 0,88$).

Analüüsime lähemalt eesti keele täis- ja kaashäälikute sagedusi tekstis, võrreldes neid andmetega soome, ungagi jt. keeltest.

V o k a a l i d järjestuvad oma sageduste põhjal eestikeelses tekstis järgmiselt (siin ja järgnevalt on võetud aluseks koondvalimi, s.o. 4 allkeele ühendteksti andmed; protsendid on arvutatud täishäälikute koguarvust tekstis, sulgudes on antud lühikeste/pikkade täishäälikute protsent):

a	26,8	(25,7/1,1)
e	24,2	(22,2/2,0)
i	20,8	(19,6/1,2)
u	13,2	(12,2/1,0)
o	6,8	(5,4/1,4)
ä	2,9	(2,4/0,5)
õ	2,9	(2,8/0,1)
ü	2,0	(1,7/0,3)
ö	0,4	(0,0/0,4)
kokku	100,0 %	(92,0/8,0)

Soome keeles on 8 täishäälikut (õ puudub). Sagedamad täishäälikud on a (24,1 % täishäälikute koguarvust tekstis), i (21,2 %), e (19,5 %), ä (12,2 %) ja o (10,4 %). Järgnevad u (9,5 %), ü (2,5 %, ortograafias y) ja ö (0,3 %). Kahe keele võrdlemisel paistab silma, et eesti keeles esineb rohkem e-, a- ja u-häälikut, kuna aga soome keeles on oluliselt rohkem o- ja eriti ä-häälikut (viimast esineb soomekeelses tekstis 12,2 % eesti 2,9 % vastu).

Võrdluseks toome andmed ka ungari keele täishäälikute sageduse kohta tekstis (Kálmán, 1963): e (ë) - 26,0 %, a - 23,6 %, o+ó - 12,5 %, i+í - 10,7 %, á - 8,4 %, é - 8,4 %, ö+ő - 4,9 %, u+ú - 3,6 %, ü+ű - 1,9 % (kokku 100,0 %).

Pikkadest täishäälikutest on eesti keeles esikohal pikk e (kirjutatud ee), mis moodustab 2,0 % vokaalide koguesinemusest tekstis (näit. sagedates sõnades see, veel, mees, een, seen, teeb). Järgnevad oo (1,4 %), ii (1,2 %), aa (1,1 %) ja uu (1,0 %). Pikka varianti kohtab kõige rohkem õ-hääliku juures (töö, öö, sööma, mööda jne.), kuna aga õ esineb peamiselt lühikesena (sõber, lõbus, õnn, või

jt.). Soome keele sagedamad pikad täishäälikud on aa (2,0 % kõigist vokaalidest tekstis), ii (1,5 %), ee (1,3 %), ää (1,2 %), uu (0,9 %).

Rühmitades eesti keele täishäälikud moodustuskoha järgi ees-, kesk- ja tagavokaalideks, saame statistilise uurimise alusel järgmised tulemused.

E e s v o k a a l i d:		
labialiseerimata	i, e, ä	47,9 %
labialiseeritud	ü, ö	2,4 %
	kokku	50,3 %
T a g a- ja k e s k v o k a a l i d:		
velaarsed	u, o, a	46,8 %
velaarpalataalne	õ	2,9 %+
	kokku	49,7 %

Taga- (ja kesk-) ning eesvokaalide suhe on eesti keeles ligikaudu 50:50. Soome keeles on taga- ja eesvokaalide suhe V. Setälä järgi 44:56 (L. Hakulineni andmetel 43:57). Ungari keeles on suhe 48:52 (Kálmán, 1963). Kui eesti keeles jätta arvestusest välja õ-häälik (keskvokaal), siis on ka eesti keeles taga- ja eesvokaalide suhe 48:52. Nimetatud kolmes keeles on eesvokaalide osatähtsus küllaltki suur. Võrdluseks võib tuua ungari naaberkeele - slovaki keele (Zsilka, 1971, 111), kus tagavokaale on 57 % eesvokaalide 43 % vastu. Leidub keeli, kus tagavokaalid on suures ülekaalus (näit. sanskritis on suhe 80:20).

K o n s o n a n t i d e sagedused on meie katse andmeil eesti keeles järgmised (protsendid on arvestatud kaashäälikute koguarvust tekstis):

+ Täpsemalt võttes on ainult lühike õ suhteliselt keskvokaal; pikk õ moodustatakse u keeleasendiga ja on seega tagapoolsem kui o ja a (T.-R. Viitso märkus). Pikk õ esineb tekstis aga suhteliselt harva (alla 0,1 %).

t d	13,4 8,4	21,8
s		16,5
k g	9,7 3,7	13,4
l		11,4
n ŋ	7,9 0,6	8,5
m		7,3
r		5,3
p b	3,5 1,3	4,8
v		4,2
j		3,5
h		3,1
võõr-konsonandid		0,2
	Kokku	100,0 %

Soome keeles on kaashäälikuid rohkem kui eesti keeles, nimelt lisanduvad heliline d ja geminaat ŋ (mida märgitakse ng-ga). Peale selle võivad samuti nagu eesti keeleski esineda mõned võõrkonsonandid (f, b, g jt.). Soome kaashäälikute sagedused tekstis on V. Setälä (1972, 38) andmete alusel järgmised: n – 19,7 %, t – 17,2 %, s – 13,5 %, k – 10,0 %, l – 9,3 %, m – 6,4 %, j – 5,4 %, h – 5,4 %, v – 4,6 %, p – 3,1 %, r – 2,9 %, d – 1,7 % ja ŋ – 0,8 % (konsonantide üldarvust). Eesti keelega võrreldes esineb soome keeles märgatavalt rohkem n-häälikut (peamiselt sõnavormi lõpus), samuti h-häälikut (eeskätt algushäälikuna).

Pikkadest kaashäälikutest, mida kirjutatakse kahe tähega, on eesti keeles esikohal ll sagedusega 1,0 % kõigist konsonantidest tekstis (näit. sagedates sõnavormides *selle*, *mille*, *teile*, *mulle*, *jälle*, *olla*, *küll*, *all*). Järgnevad nn (0,5 %), tt (0,4 %), kk (0,4 %), ss (0,3 %), mm (0,2 %), rr (0,1 %), pp (0,1 %). Eesti keeles võivad esineda ka vv ja hh, kuid nende esinemus kõnes on tühine. Võrdluseks toome sagedamad kahe tähega kirjutatud konsonandid soome keeles: ll (1,5 %), tt (1,5 %), nn (0,5 %), ss (0,5 %). Nagu varem juba nimetatud, kuuluvad pikkade kaashäälikute hulka eesti keeles ka mõned sõnasisesed ja -lõ-

pulised klusiilid, mida märgitakse ühe tähega (näit. k sõnades pika, kompvek).

M o o d u s t u s v i i s i järgi saame järgmised eesti kaashäälikute rühmad koos esinemissagedusega tekstis:

klusiilid	k/G, p/B, t/D	40,0 %
spirandid	j, s, l, r, v, h	
(kaasa arvatud võõrhäälikud		
f ja š)		44,2 %
nasaalid n/ŋ, m		15,8 %
	kokku	100,0 %

Soomekeelses tekstis esineb klusiile 31,7 %, spirante 39,5 % ja nasaale 26,8 %. Järelikult on klusiilide ja spirantide osatähtsus eestikeelses kõnes oluliselt suurem kui soome keeles, kuna aga nasaalide poolest on rikkam soome keel (tegelikult n-hääliku arvel, sest m-häälik esineb eesti keeles pisut sagedamini) Spirantidest esinevad eesti keeles oluliselt sagedamini s, l ja r, kuna aga soome keeles on sagedamad h, v ja j.

M o o d u s t u s k o h a järgi liigitatakse eesti konsonante järgmiselt (Ariste, 1968, 90-91):

labiaalid p/B, m, v, (f)	16,5 %
alveodentaalid t/D, n	29,7 %
velaarpalataalid k/G, ŋ, s,	
j, l, r, (š), (z), (ž)	50,7 %
larüngaal h	3,1 %
kokku	100,0 %

Arvestades keeletipu (apex) ja keeleselja (dorsum) aktiivsust alveodentaalide ja velaarpalataalide moodustamisel, võime nimetatud konsonandid omakorda jagada järgmisteks rühmadeks:

predorsaal-apikaalid t/D, n, s, (z), l, r	62,9 %
mediodorsaalid j, (š), (ž)	3,5 %
postdorsaalid k/G, ŋ	14,0 %
kokku	80,4 %

(Ülejäänud 19,6 % on labiaalid ja larüngaalid.)

10

Paistab silma keeletipu- ja eeskeelehäälikute (predorsaal-apikaalide) suur osatähtsus eestikeelses kõnes (62,9% kõigist konsonantidest ja 34,3 % kõigist häälikutest). Nähtus on omane väga paljudele keeltele, nii näiteks on J. Krámský (1959) poolt uuritud 23 eri keelkondadesse kuuluvas keeles predorsaal-apikaalide esinemissagedus üle 50% (kõigist konsonantidest).

Soome keeles on labiaale 14,1 %, predorsaal-apikaale 64,3 %, medio-dorsaale 5,4 %, post-dorsaale 10,8 % ja larüngaale 5,4 %. Erinevus eesti keelest ei ole kuigi suur, eriti kui ühendada tagapoolsed kaashäälikud (post-dorsaalid ja larüngaalid) ühte rühma. Eesti keeles on neid 17,1 %, soome keeles 16,2 %.

H e l i l i s e d konsonandid on eesti keeles l, m, n, r, v (välja arvatud sõna lõpus, kui neile eelneb h või s, näit. lehm, mahl, rasv; vt. Ariste, 1968, 42). Peale selle hääldub heliliselt j ja teatavatel juhtudel ka h (Ariste, 1968, 71 ja 72). Võttes arvesse ülalmainitut, võime meie materjali põhjal konstateerida, et eestikeelses tekstis (kõnes) esineb helilisi kaashäälikuid ligikaudu 41 % ja helituid 59 %. Kõigi häälikute ulatuses (kaasa arvatud vokaalid, mis on helilised), on heliliste ja helitute häälikute suhe 68:32.

Soome keeles (tekstis) on umbes 53 % helilisi kaashäälikuid (suur osakaal on n-häälikul) ja 48 % helituid. Kõigi häälikute ulatuses on heliliste ja helitute suhe 75:25. Võrdluseks võib tuua, et näit. saksa keeles moodustavad helilised konsonandid 63 % kõigist kaashäälikutest ja üldsuhe on 78:22 (Meier, 1964, 252), rootsi keeles on vastavad arvud 66 % ja 80:20 (Weiss, 1961, 51).

Eelöeldust ilmneb, et eesti keeles on mõnevõrra vähem helilisust kui mõnedes teistes keeltes. Kuid seda korvab asjaolu, et vokaalide osatähtsus on eesti keeles küllaltki suur ja kõnes esineb sageli kõlav l-häälik (eesti keeles 6,2 %, soome keeles 4,8 %, rootsi keeles 4,1 %, saksa keeles 4,0 %, vene keeles 3,5 %, läti keeles 2,5 %).

Esitame lõpuks ülevaate viiest sagedamast kaashäälikust kümnes eri keeles:

eesti keel	t/D	s	k/G	l	n
soome keel	n	t	s	k	l
ungari keel	t	l	n	m	k
mansi keel	t	l	m	χ	j
läti keel	s	r	t	k	n
leedu keel	s	k	t	r	n
vene keel	n	t	s	r	v
inglise keel	n	t	s	θ	d
saksa keel	n	r	t	d	s
rootsi keel	n	t	r	d	s

Torkab silma suur lähedus selliste sugulaskeelte vahel nagu läti ja leedu keel, saksa ja rootsi keel. Ka eesti ja soome keeles on kõik viis sagedamat kaashäälikut ühised (ka järjekord on sama, välja arvatud n-häälik). Teatavat sarnasust võib märgata ungari ja mansi keele konsonantide jaotumuses. Kõigis vaadeldavates keeltes, välja arvatud ungari ja mansi keel, on viie sagedama konsonandi seas häälikud n, s, t.

Algus- ja lõpptähtede (resp. -häälikute) sagedused tekstis ja sõnastikus. Tähestikuliste mõõdete seas on oma kindel koht tähtede (häälikute, foneemide) sagedustel p o s i t s i o o n i j ä r g i sõnas või sõnavormis. Erilist tähtsust omavad algus- ja lõpptähtede (häälikute, foneemide) sagedused, mida kasutatakse mitmesuguste lingvistilis-tüpoloogiliste ülesannete lahendamiseks (näit. Андреев, 1967; Maneca, 1967). Olulisi tulemusi on saavutatud vene ja ukraina keele morfoloogilisel automaatanalüüsimisel sõnulõpu tähtede ja tähekombinatsioonide statistika abil (Белоногов, 1971; Савченко, 1970). Huvipakkuv on G. Herdani (1963) katse kindlaks määrata inglise keele sõnavara etümoloogiliste komponentide muutumist eri perioodidel sõnade algustähtede sageduse uurimise teel. Tähtede positsioonilised sagedused tekstis ja sõnastikus on olulise tähtsusega ka mõningate informaatika-alaste probleemide uurimisel, näit. sõnade komprimeerimisel (vt. Курбаков, 1968, 163 jj.).

Eesti keele tähtede positsioonilisi sagedusi on varem uurinud H. Holm (Хольм, 1965a ja 1965б) ning Ü. Kaasik ja E. Laugaste (1969). Esitame täienduseks ja sageduste stabiilsuse kontrollimiseks käesoleva uurimistöö andmed, kusjuures toome võrdlusmaterjali soome jt. keeltest ning vaat-

leme algus- ja lõpptähtede (-häälikute) sagedusi mõnest uuest aspektist. Lisame ka andmed sõnastikusageduste kohta, mida seni pole eesti keeles uuritud. Meie koondvalim koosneb 4 allkeele osavalimitest (ilukirjandusproosa autorikõne, tegelaskõne, populaarteaduslik tekst ja ajaleht) üldmahuga 4097 sõnet. Uurimuse tulemusena saadud algus- ja lõpptähtede sagedusjärjestused on eesti keeles võrdsed vastavate h ä ä l i k u t e sagedusjärjestusega (häälikute puhul tuleb ainult ühendada k ja g, t ja d, p ja b sagedused).

Tabelis 8 esitatakse a l g u s t ä h t e d e s a g e d u s e d t e k s t i s. Kõik viis sagedamat algustähte on konsonandid: k (14,1 %), t (9,5 %), s (8,8 %), m (8,3 %), p (7,4 %). Järgnevad o, v, e, j, a, n, l, h, r, i, ü, õ, u, ä, d, f, b, ö. Viis sagedamat algustähte katavad tekstist 48,1 % ja kümme sagedamat 78,9 %. Kokku on algustähtede hulgas 74,8 % konsonante ja 25,2 % vokaale. Ü. Kaasiku ja E. Laugaste uurimuse andmeil (ilukirjandus- ja ajalehetekstid) on viis sagedamat algustähte: k (13,2 %), t (10,7 %), s (9,2 %), p (7,8 %), v (7,5 %), mis kokku moodustavad 48,4 %. H. Holmi andmeil (publitsistlik tekst) on algustähtede järjestus: k, t, p, s, v. Kõigile nimetatud uurimustele on seega ühised k, t, s, p, kusjuures tähtedele k ja t kuuluvad kindlalt kaks esimest kohta. Meie uurimistöö andmeil on m-täht võrdlemisi suure sagedusega; see tuleneb tegelaskõne kaasatõmbamisest koondvalimisse. Tabelist 8 nähtub, et peale ühiste joonte on allkeeltel ka mõned olulised erinevused. Näiteks on tegelaskõnes esikohal m-täht, mis seletub 1. isiku asesõna ohtra kasutamisega (*mina, mind, meie* jt.). Tegelaskõnes esinevad keskmisest oluliselt sagedamini ka algustähed s ja a (sõnavormide *sina, sind, sa, ta, see, seda, aga* jt. sage kasutamine), kuna aga näiteks v-täht esineb märksa harvemini, võrreldes teiste allkeeltega. Erinevused mõnede algustähtede sageduste vahel eri allkeeltes on küllaltki suured, näit. k-tähte esineb tegelaskõnes 11,7 %, ajalehes aga 17, 3 %. Osa tähti esineb allkeelte lõikes stabiilselt, näit. t (sagedused: 8,6 - 9,6 - 9,7 - 10,0 %).

Soome keeles on V. Setälä (1972) järgi sagedamad algustähed j (13,0 %), s (10,9 %), k (9,9 %), h (9,5 %), t

(9,3 %). Konsonante on algustähtede hulgas 80 % ja vokaale 20 %. Viis sagedamat algustähte katavad tekstist 52,6 %. Võrreldes eesti keelega seisneb peamine erinevus selles, et sageli esinevad algustähtedena j ja h. Kui aga vaadelda algusvokaalide sagedusi, siis ilmneb, et sagedusjärjestus on mõlemas keeles sama : o, e, a, i.

Ungari keele publitsistlikus tekstis on kuus sagedamat algustähte k, e, s, v, m, t (Статистико-комбинаторное моделирование, 1965, 207). Ühised sagedamini esinevad algustähed (resp. -häälikud) on kolmes sugulaskeeles k, t ja s.

Võrdluseks toome andmed uurimustest teiste keelte kohta (Статистико-комбинаторное моделиров.,1965; Петрова, 1968; Meier, 1964). Läti keel: p, s, a, i: vene keel: п, с, о, т, к ; inglise keele: s, t, w, f, c; saksa keel: d, s, w, f, u; prantsuse keele: a, d, l, s, e.

L õ p p t ä h t e d e s a g e d u s e d t e k s t i s on arvutatud samade allkeelte ja koondteksti alusel, mida kasutati algustähtede uurimisel.+ Tulemused esitatakse tabelis 9. Koondtekstis on kõige sagedamad lõpptähed a (19,1 %), e (17,7 %), s (13,0 %), d (11,6 %), i (11,6 %). Järgnevad t, l, n, u, b, k, m, g, r, h, v, o, p, ö. Tähelepanuväärne on seik, et viis sagedamat lõpptähte moodustavad 73,0 % kõigist lõpptähtedest tekstis, 10 sagedamat lõpptähte aga 96,2 %. Vokaale on lõpptähtede hulgas 52,7 % ja konsonante 47,3 %. Seega võib nentida olulisi erinevusi, võrreldes algustähtede sagedustega. Eestikeelses kõnes on ülekaalus vokaallõpud, mida peetakse parema kuuldavuse tingimuseks (Maneca, 1967).

Varasemate uurimuste järgi on eestikeelses tekstis sagedamad lõpptähed a, e, s, i, d (Kaasik, Laugaste, 1969; Kaasik, Tuldava, 1979); e, a, d, s, i (Хольм, 1965). Seega langevad kõigis uurimustes viis sagedamat lõpptähte kokku, kuigi järjestus pole täpselt sama.

Tabelist 9 nähtub, et eri allkeelte osas võib kohati täheldada olulisi erinevusi, näit. on teaduslikes tekstides esikohal e sagedusega 23,9 %, kuna aga sama lõpptähe

+ Vt. ka eesti keele sõnavormide pöördsagedussõnastiku materjalide põhjal saadud andmeid autorikõne kohta (Kaasik, Tuldava, 1980).

sagedus teistes allkeeltes kõigub 15-18 % vahel (vrd. lk. 77). Tegelaskõnes esinevad a ja n lõpptähena palju sagedamini kui teistes allkeeltes.

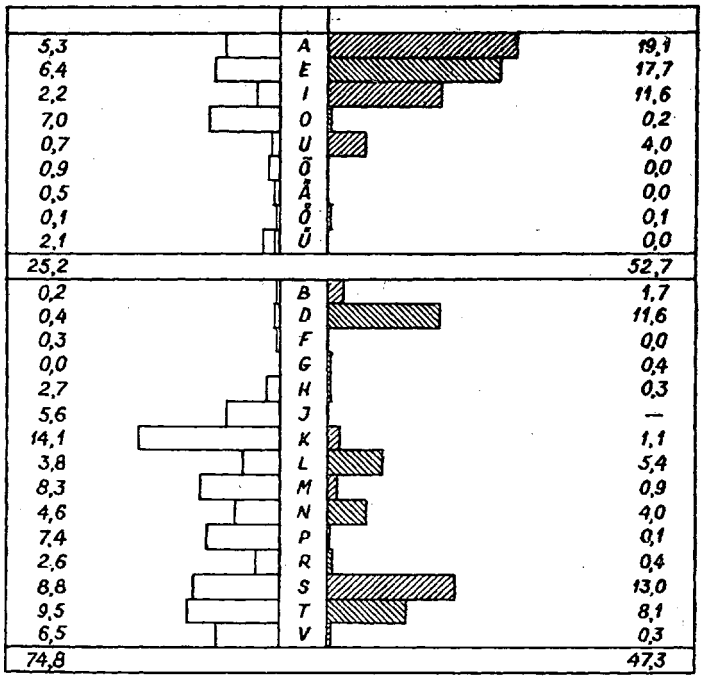
Soome keeles (Setälä, 1972, 26) on lõpptähtede sagedusjärjestus niisugune: n (28,1 %), a (23,7 %), ä (13,1 %), i (11,0 %), e (10,1 %). Viis sagedamat lõpptähte moodustavad kokku 86,0 %, seega veelgi rohkem kui eesti keeles (73,0 %). Järgnevad t, s, o, u, ö, y, r. 10 sagedamat tähte katavad 99,9 % lõpptähtede koguesinemusest. Vokaale on 60,7 % ja konsonante 39,3 %.

Nii eesti kui soome keeles ilmneb lõpptähtede kontsentratsioon, s.t. vähesed tähed (resp. häälikud) katavad suure osa kõigist lõpptähtedest (-häälikutest). Eesti keeles esinevad sõna lõpus harva niisugused tähed nagu h, v, o, p, ö ja väga harva tähed ü, ä, õ ning võõrtähed. Eesti täht j võib esineda lõpptähena ainult võõrnimedes (näit. Raj Kapoor). Lõpptähed peegeldavad grammatilisi suhteid tekstis, nii näiteks on eesti keeles d-täht mitmuse tunnuseks nii substantiivides kui verbides. Soome sagedaim lõpptäht n esineb käändelõppudes (näit. genitiivis, illatiivis), t on mitmuse tunnus.

Ungari keeles on sagedamad lõpptähed k, t, n, s. Seega domineerivad lõppasendis konsonandid. Läti keeles esineb sõna lõpus kõige sagedamini s (ülejäänud tähtede kohta puuduvad andmed). Vene keeles seisavad lõpptähtede sagedusjärjestuse alguses vokaalid: ы, е, о, а. Inglise keeles on viis sagedamat lõpptähte e, s, t, d, n.

Joonisel 4 esitatakse meie koondvalimi põhjal eesti algus- ja lõpptähtede sagedused võrdlevalt. Võrdlus näitab kujukalt, missugust osa täidab üks või teine täht eelistavalt - kas leksikaalset (sõna algul) või grammatilist osa (sõna lõpul). Paistab silma, et niisugused tähed nagu k, p, j, v, r, h, o eelistavad sõnaalgulist positsiooni, kuna aga d, u, a, e, i esinevad rohkem sõna (sõnavormi) lõpus. Tähti s ja t kasutatakse enam-vähem võrdselt nii algus- kui ka lõppasendis.

Algus- ja lõpptähtede sagedusi võib täiendada andmetega tähtede sõnasisese esinemuse kohta. Kui arvestada iga tähe üldsageduseks 100 %, siis jaotuvad näit. tähe a puhul sagedused algus-, sise- ja lõpptähena järgmiselt: 7 % -



Joon. 4. Algus- ja lõpptähed tekstis.

- 67 % - 26 %. Järelikult kasutatakse a-tähte kõige sagedamini sõnasiseses positsioonis. Alljärgnevalt esitame koondvalimi andmed kõigi tähtede kohta esinemuse kohta tekstis (arvud tähe järel tähistavad esinemissagedust protsentides vastavalt algus-, sise- ja lõpptähena):

a	7-67-26	h	27-70-3	n	16-70-14	u	2-87-11
b	5-52-43	i	4-76-20	o	33-66-1	v	48-50-2
d	1-55-44	j	52-48-0	p	67-32-1	õ	11-89-0
e	9-65-26	k	44-52-4	r	15-83-2	ä	5-95-0
f	58-42-0	l	10-76-14	s	17-59-24	ö	9-87-4
g	0-96-4	m	36-60-4	t	22-59-19	ü	35-65-0

Tekstisageduste kõrval pakuvad huvi algus- ja lõpptähtede sagedused s õ n a s t i k u s. Et saaks paremini võrrelda teksti- ja sõnastikusagedusi, võtame aluseks ühe-ainsa allkeele, nimelt ilukirjandusproosa autorikõne andmed

küllaltki suure valimi põhjal (kokku 15.000 sõnet). Materjal on võetud võrdsetes osades eri autorite teostest. Viis sagedamat algustähte on k, t, p, s, v/j nii tekstis kui sõnastikus, kusjuures tekstis on esimesed kaks alati k ja t, sõnastikus k ja p. Autorikõne sõnavormide sõnastikus on algustähtede sagedused järgmised: k (16,6 %), p (10,1 %), s (9,8 %), t (p,5 %), v (8,5 %). Vokaalid moodustavad sõnastiku algustähtede koguesinemusest 16,0 % (tekstist 20,0 %). Algustähtede sagedused tekstis ja sõnastikus on kokkuvõttes tugevas korrelatsioonis (seda on täheldatud ka teiste keelte puhul, vt. Karlgren, 1962). Mõnevõrra suurem erinevus teksti ja sõnavormide sõnastiku vahel ilmneb lõpptähtede osas. Sagedamad lõpptähed autorikõne sõnastikus on s, a, e, d, t (tekstis a, s, e, i, d), kusjuures vokaalide osakaal on 46,7 % (tekstis 53,8 %). Kokkuvõtlikud andmed teksti- ja sõnastikusageduste kohta koos tekstikoormuse (T/S-suhte) arvutamisega esitame tabelis 10.

Tähtede korrelatiivne funktsioon. Tähtede funktsionaalset koormust eri asendites väljendatakse keelestatistikas nn. k o r r e l a t i i v s e f u n k t s i o o n i näol, mida määratletakse kui "keeleelemendi tingtõenäosuse suhet sõltumatusse tõenäosusse" (Андреев, 1967, 22). Näiteks: p-tähe "sõltumatu tõenäosus" ehk teiste sõnadega suhteline sagedus tekstis üldse on 1,8 %, algustähena aga 7,4%. Viimane väljendab "tingtõenäosust" antud olukorras, s. o. algpositsioonis. Korrelatiivne funktsioon on sel juhul 7,4: :1,8 = 4,1. See tähendab, et p-täht esineb algustähena umbes neli korda sagedamini kui keskmiselt tekstis üldse, seega on p-tähe funktsionaalne koormus tekstis olulisel määral koondatud sõna algusse. Samas aga on p-tähe sagedus lõpptähena ainult 0,1 % (kõigist lõpptähtedest), s. t. korrelatiivne funktsioon KF = 0,1 : 1,8 = 0,06. Tähe funktsionaalne koormus sõna lõpul on seega väga väike.

Korrelatiivse funktsiooni väärtused kõigi tähtede kohta alg- ja lõpp-positsioonis esitatakse tabelis 11. Ilmneb, et s õ n a a l g u l i s e s p o s i t s i o o n i s on eesti keeles kõige suuremad korrelatiivse funktsiooni väärtused järgmistel tähtedel: p (4,1), j (3,1), v (3,0), k (2,7), m (2,2), ü (2,1), o (2,0). Väikesed KF väärtused on tähtedel ä ja ö (mõlemal 0,3), i (0,2) ja u (0,1). Peale

selle on KF väärtused väikesed tähtedel b, g, d, mis teatavasti võivad eesti keeles esineda sõna algul ainult võõrsõnades. Seevastu on täht f spetsiifiline algustäht (KF=3,0).

Soome keeles on suurte KF väärtustega järgmised algustähed: j (4,9), p (3,5), h (3,5), v (2,4), m (2,4), k (2,0). Võrreldes eesti keelega on sarnasus suur, erineb ainult h-tähe koormus, mis soome keeles on algustähe osas väga suur. Kui võrrelda korrelatiivse funktsiooni väärtusi kahes keeles vokaalide ja konsonantide osas, saame järgmised tulemused: eesti keeles on vokaalide korrelatiivne funktsioon 0,6, konsonantide KF aga 1,4; soome keeles on vastavad väärtused 0,4 ja 1,5. Mõlemas keeles on konsonantide funktsionaalne koormus algustähena oluliselt suurem kui vastav vokaalide koormus.

Võrdluseks toome andmed mõne teise keele kohta (Андреев, 1967). Suuremad KF väärtused on järgmistel algustähtedel: läti keeles - p, n, s; vene keeles - П, С, В; tšehhi keeles - p, v, s; saksa keeles - p, w, b, prantsuse keeles p, c, d, hispaania keeles - p, c, e, itaalia keeles - c, p, s. Kõigis nimetatud keeltes on samuti nagu eesti ja soome keeleski sõnaalgulises positsioonis suure funktsionaalse koormusega labiaal p. Labiaalid ja labiodentaalid on aktiivsed ka inglise keeles: w, b, f. Ungari keeles on suurima KF väärtusega algustähed v (6,0), k (3,1), m (2,0), s (1,7), e (1,4) ja t (0,9).

S õ n a l õ p u l i s e s p o s i t s i o o n i s on eesti keeles eriti koormatud järgmised tähed (vt. tabel 11): d (KF = 2,6), b (2,4), a (1,6), e (1,6), s (1,5), i (1,2) ja t (1,1). Ülejäänud tähtede KF väärtus on alla 1. Soome keeles saame järgmise pingerea: n (2,8), ä (2,2), a (2,0), e (1,1) ja i (1,1), t (1,0). Põhiline erinevus seisneb selles, et eesti keeles on aktiivne lõpptäht -b (esineb lõpptähena peamiselt verbi oleviku ainsuse 3. pöördes, näit. tuleb, võtab; soome keeles on vastav vorm vokaallõpuline), soome keeles aga esineb lõppasendis kõige sagedamini -n (näit. genitiivi- ja illatiivivormi lõpus; jalan, jalkaan jne.; mõningates asesõnavormides, näit. kukin, kunkin, mikin, minkin, kukaan jt.; umbisikulises tegumoes, näit. otetaan; eesti keeles on need juhud vokaallõpulised). Vokaalide korrelatiivne funktsioon sõnalõpulises positsioonis on eesti kee-

les 1,1 ja soome keeles 1,3; konsonantide KF eesti keeles 0,9 ja soome keeles 0,8.

Kui sõnaalgulises positsioonis võis täheldada küllaltki suurt sarnasust paljude keelte vahel tähtede korrelatiivse funktsiooni seisukohast, siis sõnalõpulised KF väärtused diferentseerivad keeli oluliselt. Näiteks on suure sõnalõpulise koormusega vene keeles tähed Й, Ь, Я, М, И, inglise keeles y, d, s, prantsuse keeles s, a, t, itaalia keeles o, e, i, ungari keeles k, s, t. Pole kahtlust, et l õ p p t ä h t e d e sagedused ja vastavad korrelatiivse funktsiooni väärtused võimaldavad kokkuvõttes küllaltki täpselt diferentseerida ja identifitseerida keeli tekstide automaattöötlemisel.

Tähtede korrelatiivse funktsiooni võib arvutada ka kõigi ülejäänud positsioonide kohta. Tabelis 12 esitame andmed sõna algusest t e i s e l k o h a l asuvate tähtede sageduste ja vastavate KF väärtuste kohta. Tulemused on küllaltki huvitavad puhtlingvistilisest seisukohast. Nimelt selgub, et KF väärtused eristavad selgepiiriliselt vokaale konsonantidest. Kui KF on suurem kui 1, siis on tegemist vokaaliga. Kõigil konsonantidel on KF väärtus alla 1, kusjuures sonoorsed r, n ja l seisavad vokaalidele kõige lähemal. Siin tutvustatud katse tulemused on kujukaks näiteks selle kohta, kuidas formaalne kvantitatiivne analüüs suudab esile tuua tähtsad kvalitatiivsed suhted keeleelementide vahel.

Tabel 1

Tähtede esinemissagedus eestikeelses tekstis
(5 allkeelt ja koondtekst; valimi üldmaht N = 78900 tähte)

Täht	Iluk. proosa	Kõne-keel	Aja-leht	Tead.-tehn.	Luule	KOKKU	
						arv	%
a	13,1	13,0	11,4	10,2	12,4	9660	12,2
e	10,0	11,6	10,5	12,7	9,2	8593	10,9
i	9,6	10,0	10,3	9,8	8,7	7757	9,8
s	8,6	8,2	8,9	9,0	8,1	6785	8,6
t	6,8	6,8	7,9	7,4	6,0	5567	7,1
l	6,7	6,5	6,4	6,1	7,0	5124	6,5
u	5,9	5,3	6,0	6,0	6,7	4603	5,8
k	5,2	4,9	5,0	4,8	5,9	3990	5,1
n	4,7	5,2	4,9	4,5	5,8	3888	4,9
m	3,7	4,9	3,7	4,1	4,0	3216	4,1
d	3,8	3,6	4,2	4,2	4,2	3081	3,9
o	3,3	3,4	4,1	4,2	4,1	2913	3,7
r	2,9	2,4	2,7	3,3	2,5	2170	2,8
v	2,3	2,0	2,3	2,3	2,4	1771	2,3
g	1,9	2,0	1,7	2,0	1,7	1507	1,9
h	2,0	1,9	1,4	1,7	2,2	1439	1,8
j	2,1	1,7	1,7	1,4	2,0	1420	1,8
p	2,0	1,6	1,7	1,7	1,8	1402	1,8
ä	1,7	1,7	1,2	1,2	1,6	1181	1,5
õ	1,5	1,1	1,1	1,4	1,2	990	1,3
ü	0,9	1,0	0,8	1,0	1,0	720	0,9
b	0,8	0,8	0,8	0,7	1,1	666	0,8
ö	0,4	0,3	0,4	0,2	0,4	254	0,3
võõr-tähed	0,1	0,1	0,9	0,1	0,0	203	0,2
Kokku (%)	100,0	100,0	100,0	100,0	100,0	-	100,0
Tähti	25420	20914	16208	11358	5000	78900	-
Allk. osa	32%	27%	20%	15%	6%	100%	-

T a b e l 2

Tähtede esinemissagedus ja hajuvushinnangud koondteksti andmete põhjal: $\bar{p}$ (%) - keskmine suhteline sagedus, s - standardhälve, $\varepsilon_{\bar{p}}$ - keskväärtuse piirviga ja $\delta_{\bar{p}}$ - suhteline viga 95%-lisel usaldusnivool

Täht	$\bar{p}$ %	s	$\varepsilon_{\bar{p}}$	Usaldus-piirid (%)	$\delta_{\bar{p}}$ %
a	12,2	1,07	1,3	10,9 ... 13,5	10,7
e	10,9	1,04	1,3	9,6 ... 12,1	11,9
i	9,8	0,39	0,5	9,3 ... 10,3	5,1
s	8,6	0,32	0,4	8,2 ... 9,0	4,7
t	7,1	0,47	0,5	6,6 ... 7,6	7,0
l	6,5	0,23	0,3	6,2 ... 6,8	4,6
u	5,8	0,37	0,5	5,3 ... 6,3	8,6
k	5,1	0,23	0,3	4,8 ... 5,4	5,9
n	4,9	0,37	0,5	4,4 ... 5,4	10,2
m	4,1	0,50	0,6	3,5 ... 4,7	14,6
d	3,9	0,25	0,3	3,6 ... 4,2	7,7
o	3,7	0,39	0,5	3,2 ... 4,2	13,5
r	2,8	0,30	0,4	2,4 ... 3,2	14,3
v	2,3	0,15	0,2	2,1 ... 2,4	8,7
g	1,9	0,12	0,2	1,7 ... 2,1	10,5
h	1,8	0,25	0,3	1,5 ... 2,1	16,7
j	1,8	0,25	0,3	1,5 ... 2,1	16,7
p	1,8	0,16	0,2	1,6 ... 2,0	11,1
ä	1,5	0,24	0,3	1,2 ... 1,8	20,0
õ	1,3	0,18	0,2	1,1 ... 1,5	15,4
ü	0,9	0,08	0,1	0,8 ... 1,0	11,1
b	0,8	0,04	0,1	0,7 ... 0,9	12,5
ö	0,3	0,03	0,1	0,2 ... 0,4	33,3

Tabel 3

Tähtede esinemissagedus ilukirjandusproosa autorikõnes ja tegelaskõnes (tärnikesega on märgitud statistiliselt olulised vahed olulisusnivool $\alpha < 0{,}001$)

Täht	Autorikõne		Tegelaskõne		Vahe
	p_a %	astak	p_t %	astak	$p_a - p_t$
a	12,8	(1)	13,0	(1)	-0,2
e	10,0	(2)	11,6	(2)	-1,6*
i	9,4	(3)	10,0	(3)	-0,6
s	8,8	(4)	8,2	(4)	+0,6
t	6,7	(6)	6,8	(5)	-0,1
l	6,8	(5)	6,5	(6)	+0,3
u	6,3	(7)	5,3	(7)	+1,0*
k	5,2	(8)	4,9	(9,5)	+0,3
n	4,3	(9)	5,2	(8)	-0,9*
m	3,4	(11)	4,9	(9,5)	-1,5*
d	4,1	(10)	3,6	(11)	+0,5
o	3,3	(12)	3,4	(12)	-0,1
r	3,0	(13)	2,4	(13)	-0,6
v	2,3	(14)	2,0	(14,5)	+0,3
g	1,9	(17,5)	2,0	(14,5)	-0,1
h	1,9	(17,5)	1,9	(16)	0
j	2,0	(16)	1,7	(17,5)	+0,3
p	2,2	(15)	1,6	(19)	+0,6*
ä	1,7	(19)	1,7	(17,5)	0
õ	1,5	(20)	1,1	(20)	+0,4
ü	0,9	(21,5)	1,0	(21)	-0,1
b	0,9	(21,5)	0,8	(22)	+0,1
ö	0,5	(23)	0,3	(23)	+0,2
Võõr-tähed	0,1	(24)	0,1	(24)	0
Kokku (%)	100,0	-	100,0	-	-
Tähti	10420	-	20914	-	-

Tabel 4

Tähtede esinemissagedus eestikeelsetes tekstides: I – "Eesti Postimees" 19.saj., vana kirjaviis (Kaasik, Laugaste, 1969); II – rahvalaul (Kaasik, Laugaste, Ääremaa, 1975), III – segatekst (Hansen, 1961); IV – ajaleht 52%, ilukirjandus 48% (Kaasik, Laugaste, Ääremaa, 1975); V – ilukirjandusproosa 32%, tegelaskõne 27%, ajaleht 20%, teaduslik-tehniline tekst 15%, luule 6%; VI – IV+V: ilukirjandus 56%, ajaleht, teaduslik-tehniline tekst – 44%.

Täht	I	II	III	IV	V	VI
a	13,8	13,7	13,0	12,9	12,2	12,6
b	0,9	0,7	0,5	0,9	0,8	0,8
d	4,3	3,8	4,3	4,0	3,9	4,0
e	11,3	12,1	11,0	11,1	10,9	11,0
g	1,7	1,2	1,6	2,2	1,9	2,1
h	2,1	1,8	1,6	1,9	1,8	1,9
i	8,6	9,0	10,0	9,4	9,8	9,6
j	2,3	1,2	1,6	1,9	1,8	1,8
k	5,0	5,5	5,7	4,9	5,1	5,0
l	5,6	7,1	6,2	6,1	6,5	6,3
m	4,3	3,9	4,1	3,8	4,1	3,9
n	4,3	5,1	4,9	4,7	4,9	4,8
o	2,8	2,8	3,1	3,5	3,7	3,6
p	1,7	2,2	1,7	1,9	1,8	1,9
r	2,3	2,6	2,7	2,7	2,8	2,7
s	8,7	8,7	8,7	8,7	8,6	8,7
t	6,9	5,7	7,0	7,1	7,1	7,1
u	6,0	5,4	5,4	6,0	5,8	5,9
v	2,5	2,4	2,4	2,2	2,3	2,2
õ	1,4	1,3	1,4	1,3	1,3	1,3
ä	2,3	2,7	1,6	1,4	1,5	1,4
ö	0,3	0,3	0,4	0,5	0,3	0,4
ü	0,9	1,0	0,9	0,8	0,9	0,8
Võõr-tähed	–	–	0,2	0,1	0,2	0,2
Kokku (%)	99,9	100,2	100,0	100,0	100,0	100,0
Tähti	43250	47700	?	84440	78900	163340

T a b e l 5

Tähtede suhtelised sagedused tekstis ja sõnastikus (sõnavormid) eesti ilukirjandusproosa autorikõne põhjal

Täht	Suhteline sagedus (%)		Vahe T-S
	Tekstis (T)	Sõnastikus (S)	
a	12,8	12,7	+0,1
e	10,0	9,0	+1,0
i	9,4	8,5	+0,9
s	8,8	10,0	-1,2
l	6,8	6,5	+0,3
t	6,7	7,1	-0,4
u	6,3	7,1	-0,8
k	5,2	5,7	-0,5
n	4,3	4,2	+0,1
d	4,1	4,3	-0,2
m	3,4	3,3	+0,1
o	3,3	2,8	+0,5
r	3,0	3,5	-0,5
v	2,3	2,4	-0,1
p	2,2	2,4	-0,2
j	2,0	1,4	+0,6
g	1,9	1,9	0
h	1,9	1,8	+0,1
ä	1,7	1,4	+0,3
õ	1,5	1,8	-0,3
b	0,9	0,8	+0,1
ü	0,9	0,9	0
ö	0,5	0,4	+0,1
Võõr-tähed	0,1	0,1	0
Kokku	100,0	100,0	-
Tähti	10420	19770	-

Tabel 6

Eesti keele häälikute sagedused tekstis: I - Saareste, 1932; II - Särgava, 1967 (N = 30.000 häälikut, ajalehed ja ajakirjad); III - Tuldava (N = 23.560 häälikut, 4 allkeelt võrdsetes osades: ilukirjandusproosa autorikõne, tegelaskõne, ajalehetekst, teadusalane tekst)

Häälik	Sagedused protsentides		
	I	II	III
Vokaalid:			
a	14,0	12-14	12,2
e	10,5	11-12	11,0
i	8,5	9	9,5
o		4-5	3,1
u		5	6,0
õ	1,5	1,5	1,3
ä		1,5-2	1,3
ö	0,2	0,2-0,5	0,2
ü	0,8	0,8-1	0,9
Konsonandid:			
p+B		3	2,6 (1,9+0,7)
t+D	10,5	10,5-11	11,9 (7,3+4,6)
k+G		8,9	7,3 (5,3+2,0)
h	1,0	1-1,5	1,7
j	2,0	1,5-2	1,9
l	7,0	7-7,5	6,2
m		3-4	4,0
n+ŋ		4-5	4,6 (4,3+0,3)
r		2-3	2,9
s	10,0	10	9,0
v		2-3	2,4
-	-	100	100,0

Tabel 7

Häälikute sagedus eestikeelses tekstis (4 allkeelt võrdsetes osades: ilukirjandusproosa autorikõne, tegelaskõne, ajalehetekst, teadusalane tekst), valimi üldmaht N = 23560 häälikut

Hääl	Suhteline sagedus (%)			Ühe ja kahe tähega kirjutatud häälikute suhe (%)
	Kokku	Ühe tähega kirjutatud	Kahe tähega kirjutatud	
a	12,2	11,7	0,5	96:4
t (11,9)	7,3	7,1	0,2	97:3
D (11,9)	4,6	4,6	-	100:0
e	11,0	10,1	0,9	92:8
i	9,5	9,0	0,5	94:6
s	9,0	8,8	0,2	98:2
k (7,3)	5,3	5,1	0,2	96:4
G (7,3)	2,0	2,0	-	100:0
l	6,2	5,7	0,5	91:9
u	6,0	5,5	0,5	93:7
n (4,6)	4,3	4,0	0,3	93:7
ŋ (4,6)	0,3	0,3	-	100:0
m	4,0	3,9	0,1	98:2
o	3,1	2,5	0,6	80:20
r	2,9	2,8	0,1	97:3
p (2,6)	1,9	1,8	0,1	96:4
B (2,6)	0,7	0,7	-	100:0
v	2,3	2,3	0,0	100:0
j	1,9	1,9	0,0	100:0
h	1,7	1,7	0,0	100:0
ä	1,3	1,1	0,2	83:17
õ	1,3	1,3	0,0	100:0
ü	0,9	0,7	0,2	85:15
ö	0,2	0,0	0,2	0:100
Võõrhäälikud	0,1	-	-	-
Kokku	100,0	94,6	5,3	95:5

T a b e l 8

Algustähtede sagedused tekstis

Algustäht	Autori-kõne	Tegelas-kõne	Tead. tekst	Aja-leht	Koondtekst arv	Koondtekst %
a	4,3	6,7	5,0	4,9	218	5,3
b	0,4	0,1	0,1	0,2	8	0,2
c	0,0	0,0	0,0	0,0	0	0,0
d	0,0	0,2	1,1	0,3	15	0,4
e	5,4	7,2	6,9	5,9	262	6,4
f	0,1	0,1	0,6	0,5	11	0,3
g	0,0	0,0	0,0	0,0	0	0,0
h	4,3	2,4	2,2	1,5	110	2,7
i	1,4	2,3	2,6	2,9	91	2,2
j	6,5	5,3	4,2	6,3	229	5,6
k	13,7	11,7	15,1	17,3	580	14,1
l	4,7	3,3	3,9	3,1	154	3,8
m	6,7	12,5	7,3	5,6	342	8,3
n	5,6	4,1	4,4	4,2	189	4,6
o	4,9	7,4	7,6	8,6	287	7,0
p	8,5	6,7	6,7	7,9	305	7,4
r	2,3	2,3	3,7	2,3	106	2,6
s	8,2	10,3	8,1	8,2	361	8,8
t	10,0	9,7	8,6	9,6	390	9,5
u	0,9	0,3	0,8	0,7	27	0,7
v	8,1	4,7	7,3	6,2	267	6,5
õ	0,8	0,6	1,2	0,9	35	0,9
ä	0,7	0,8	0,2	0,0	20	0,5
ö	0,3	0,1	0,0	0,2	6	0,1
ü	2,2	1,2	2,4	2,7	84	2,1
Kokku (%)	100,0	100,0	100,0	100,0	-	100,0
Algus-tähti	1106	1230	900	861	4097	-

Tabel 9

Lõpptähtede sagedused tekstis

Lõpp-täht	Autori-kõne	Tegelas-kõne	Tead. tekst	Aja-leht	Koondtekst arv	Koondtekst %
a	18,2	23,8	14,6	18,1	805	19,1
b	1,6	1,7	2,1	1,4	72	1,7
d	12,6	9,6	12,1	12,7	489	11,6
e	14,2	15,5	23,9	18,4	745	17,7
g	0,6	0,2	0,6	0,5	19	0,4
h	0,1	0,9	0,0	0,2	14	0,3
i	12,1	12,1	10,3	11,5	488	11,6
k	1,3	0,9	1,3	0,9	46	1,1
l	5,4	4,9	5,5	6,1	227	5,4
m	1,2	1,1	0,7	0,5	38	0,9
n	4,1	5,9	3,2	2,2	171	4,0
o	0,1	0,3	0,1	0,5	10	0,2
p	0,0	0,3	0,0	0,0	4	0,1
r	0,5	0,0	0,3	0,7	15	0,4
s	14,5	12,7	11,0	13,8	550	13,0
t	8,5	6,2	9,5	8,9	342	8,1
u	4,6	3,4	4,5	3,4	168	4,0
v	0,2	0,3	0,3	0,2	11	0,3
ä	0,0	0,1	0,0	0,0	1	0,0
ö	0,2	0,1	0,0	0,0	3	0,1
Kokku (%)	100,0	100,0	100,0	100,0	-	100,0
Tähti	1115	1276	953	874	4218	-

Tabel 10

Algus- ja lõpptähtede sagedused ilukirjandusproosa autorikõne tekstis ja sõnavormide sõnastikus koondvalimi põhjal

Täht	Algustähena			Lõpptähena		
	tekstis (T)	sõnastikus (S)	T/S	tekstis (T)	sõnastikus (S)	T/S
a	4,2	4,2	1,0	22,4	16,1	1,4
b	0,7	0,2	3,5	0,8	1,1	0,7
d	0,2	0,2	1,0	11,0	13,8	0,8
e	4,8	3,0	1,6	13,2	15,6	0,9
g	0,1	0,1	1,0	0,9	0,4	2,3
h	3,3	4,6	0,7	0,1	0,1	1,0
i	1,9	1,8	1,2	12,8	9,8	1,3
j	7,5	3,5	2,1	-	-	-
k	14,7	16,6	0,9	0,9	1,2	0,8
l	4,7	6,7	0,7	5,1	5,3	0,9
m	5,7	6,1	0,9	0,7	0,8	0,9
n	5,2	4,2	1,2	0,9	0,5	1,8
o	5,1	2,0	2,6	0,2	0,2	1,0
p	8,7	10,1	0,9	0,1	0,1	1,0
r	2,8	3,8	0,7	0,8	0,7	1,1
s	8,9	9,8	0,9	15,8	17,4	0,9
t	10,2	9,5	1,1	9,5	11,3	0,8
u	0,8	1,1	0,7	4,4	4,9	0,9
v	7,2	8,5	0,9	0,3	0,6	0,5
õ	0,9	1,4	0,6	0,0	0,0	0,0
ä	0,5	0,6	0,8	0,0	0,0	0,0
ö	0,2	0,2	1,0	0,1	0,1	1,0
ü	1,6	1,7	0,9	0,0	0,0	0,0
Võõrtähed	0,1	0,1	1,0	0,0	0,0	0,0
Kokku (%)	100,0	100,0	-	100,0	100,0	-
Sõnesid, sõnavorme	14981	8192	-	14981	8192	-

Tabel 11

Algus- ja lõpptähtede suhtelised sagedused tekstis ja korrelatiivne funktsioon (KF) 4 allkeele koondteksti põhjal

Täht	Tekstis üldse (T)	Algus-tähena (A)	KF = A/T	Lõpp-tähena (L)	KF = L/T
a	12,0	5,3	0,44	19,1	1,59
b	0,7	0,2	0,29	1,7	2,43
d	4,4	0,4	0,09	11,6	2,64
e	11,3	6,4	0,57	17,7	1,56
f	0,1	0,3	3,00	0,0	-
g	1,9	0,0	-	0,4	0,21
h	1,7	2,7	1,59	0,3	0,18
i	9,5	2,2	0,23	11,6	1,22
j	1,8	5,6	3,11	-	-
k	5,3	14,1	2,66	1,1	0,21
l	6,4	3,8	0,59	5,4	0,84
m	3,8	8,3	2,18	0,9	0,24
n	4,7	4,6	0,96	4,0	0,85
o	3,5	7,0	2,00	0,2	0,06
p	1,8	7,4	4,11	0,1	0,06
r	2,8	2,6	0,93	0,4	0,14
s	8,8	8,8	1,00	13,0	1,49
t	7,1	9,5	1,34	8,1	1,14
u	6,1	0,7	0,11	4,0	0,65
v	2,2	6,5	2,95	0,3	0,14
õ	1,3	0,9	0,69	0,0	-
ä	1,5	0,5	0,33	0,0	-
ö	0,3	0,1	0,33	0,1	0,33
ü	1,0	2,1	2,10	0,1	-
Kokku (%)	100,0	100,0	-	100,0	-
Tähti	24796	4097	-	4097	-

Tabel 12

Algustähele järgneva tähe suhteline sagedus tekstis ja korrelatiivne funktsioon (KF) 4 allkeele koondvalimi põhjal

Täht	Tekstis üldse (T)	Teisel kohal (T_2)	KF = T_2/T
õ	1,3	6,2	4,8
ä	1,5	6,0	4,0
ü	1,0	2,3	2,3
š	0,3	0,6	2,0
o	3,5	6,4	1,8
a	12,0	18,8	1,6
u	6,1	9,6	1,6
e	11,3	15,4	1,4
i	9,5	11,9	1,3
r	2,8	2,3	0,9
n	4,7	4,1	0,9
l	6,4	5,2	0,8
g	1,9	1,4	0,7
h	1,7	1,0	0,6
t	7,1	2,2	0,3
b	0,7	0,2	0,3
s	8,8	2,5	0,3
j	1,8	0,5	0,3
k	5,3	1,4	0,3
p	1,8	0,4	0,2
m	3,8	0,8	0,2
d	4,4	0,4	0,1
v	2,2	0,1	0,05
Kokku (%)	100,0	100,0	-
Tähti	24796	4097	-

VIITEKIRJANDUS

Ariste, P. Eesti keele foneetika. 3. trükk. – Tartu: TRÜ, 1968.

Arndorfer, H. Das Phänomen der Lautbedeutsamkeit aus biophonetischer Sicht. – Proceedings of the Sixth International Congress of Phonetic Sciences. Prague, 1967, 125–127.

Bourdon, B. L'expression des émotions et des tendences dans le langage. Paris, 1892.

Ertel, S. Psychophonetik. Untersuchungen über Lautsymbolik und Motivation.–Göttingen: Hogrefe, 1969.

Guiraud, P. Les caractères statistiques du vocabulaire. Essai de méthodologie. Paris, 1954.

Hakulinen, L. Suomen kielen rakenne ja kehitys. Helsinki, 1941.

Hansen, H. Ladumismasinamatriitside ratsionaalsem komplekteerimine. – Polügrafist , 1961, nr. 2, 21–24.

Herdan, G. The Calculus of Linguistic Observations. – The Hague: Mouton, 1962.

Herdan, G. A Method for the Quantitative Analysis of Language Mixture. – Statistical Methods in Linguistics, 2.–Stockholm: Skriptor, 1963, 110–123.

Kaasik, Ü., Laugaste, E. Tähtede sagedus eestikeelsetes tekstides. – Keel ja Kirjandus , 1969, nr. 10, 600–605.

Kaasik, Ü., Laugaste, E., Ääremaa, K. Tähtede ja silpide sagedus eestikeelsetes tekstides. – Keel ja Kirjandus , 1975, nr. 1, 21–29.

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

Kaasik, Ü., Tuldava, J., Villup, A., Ääremaa, K. Eesti tänapäeva ilukirjandusproosa autorikõne tekstide sagedussõnastik. – Töid keelestatistika alalt, II. TRÜ Toimetised, vihik 413. Tartu, 1977, 5–140.

Kaasik, Ü., Tuldava, J. Sõnalõpu ja sõnapikkuse vahekorrast eestikeelses tekstis. – Eesti keele sõnavarastatistika erikäsimusi. Töid keelestatistika alalt, IV. TRÜ Toimetised, vihik 477. Tartu, 1980, 154–167.

Kaasik, Ü., Tuldava, J., Ääremaa, K. Eesti keele sõnavormide pöördsagedussõnastik. – Eesti keele sõnavarastatistika erikäsimusi. Töid keelestatistika alalt, IV. TRÜ Toimetised, vihik 477. Tartu, 1980, 5–153.

Kálmán, B. Häälikustatistilisi vaatlusi mansi keele alalt. – Nonaginta. J.V. Veski 90. sünnipäevaks 27. juunil 1963. Emakeele Seltsi Toimetised nr. 6. Tallinn, 1963, 123–127.

Karlgren, H. Statistical methods in phonetics. – Manual of Phonetics. Edited by B. Malmberg. – The Hague: North-Holland Publishing Co., 1968, 129–154.

Kask, A. Eesti keele ajalooline grammatika, I. Häälikulugu. 2. trükk.–Tartu: TRÜ, 1972.

Krámský, J. Fonologické využití samohláskových fonémat. - Linguistica Slovaca, IV - VI. Bratislava, 1948.

Krámský, J. A Quantitative Typology of Languages. - Language and Speech. Vol. 2, 1959.

Kučera, H., Monroe, G.K. A Comparative Phonology of Russian, Czech and German. New York, 1968.

Lepamaa, A. Tõenäosusteooria ja matemaatiline statistika. Tallinn, 1968.

Mackay, A. On the Type-Fount of the Phaistos Disc. - Statistical Methods in Linguistics, 4.-Stockholm; Skriptor, 1965, 15-25.

Maneca, C. Considérations statistiques sur les finales vocaliques en roumain. - Proceedings of the Sixth International Congress of Phonetic Sciences. Prague, 1967, 609-611.

Marchand, H. Phonetic symbolism in English Word-Formation. - Indogermanische Forschungen. Bd. 64, H.2. Berlin, 1959.

Martinet, A. Economie des changements phonétiques. Berne, 1955.

Meier, H. Deutsche Sprachstatistik, I - II. Hildesheim, 1964.

Moreau, R. Quantité d'information et redondance aux différents niveaux du français écrit. Paris, 1963.

Mustonen, S. Multiple Discriminant Analysis in Linguistic Problems. - Statistical Methods in Linguistics, 4. - Stockholm; Skriptor, 1965, 37-44.

Mäger, M. Luuleteksti fonaetilise struktuuri ja semantika vahekorrast. - Keel ja Kirjandus , 1971, nr. 6, 325-334.

Pierce, J.R. Symbols, Signals and Noise: The Nature and Process of Communication.-London: Hutchinson, 1962.

Saareste, A. Die estnische Sprache. Tartu, 1932.

Setälä, V. Suomen kielen dynamiikkaa.-Helsinki: Suomalaisen Kirjallisuuden Seura, 1972.

Sigurd, B. A Note on the Number of Phonemes. - Statistical Methods in Linguistics, 2.-Stockholm; Skriptor, 1963, 94-99.

Särgava, V. Eestikeelse kõne kasutamine kuulmise uurimisel. - Keel ja Kirjandus , 1967, nr. 1, 27-30.

Trubetzkoy, N.S. Grundzüge der Phonologie. - Travaux du Cercle Linguistique de Prague, VII. Praha, 1939.

Tuldava, J. Statistiline väljavõttemeetod keeleteaduses. - Linguistica, I.-Tartu; TRÜ, 1969, 5-49.

Tuldava, J. Informatsiooniteooria ja keeleteadus. - Keel ja Kirjandus , 1970a, nr. 6, 329-339.

Tuldava, J. Statistilised testid keeleteaduses. - Linguistica, II.-Tartu; TRÜ, 1970b, 125-196.

Tuldava, J. Sagedussõnastik leksikostatistilise uurimise objektina. - Töid keelestatistika alalt, II. TRÜ Toimetised, vihik 413. Tartu, 1977, 141-171.

Tuldava, J. Sõnavormide esinemus eestikeelses tekstis. - Eesti keele sõnavarastatistika probleeme. Töid keelestatistika alalt, III. TRÜ Toimetised, vihik 446. Tartu, 1978, 107-126.

Verglas, A. Remarques sur la relation entre rang et fréquence des lettres dans les mots français. - Bulletin d'information du laboratoire d'analyse lexicologique. Vol. 6. Besançon, 1963.

Weiss, M. Über die relative Häufigkeit der Phoneme des Schwedischen. - Statistical Methods in Linguistics, I.-Stockholm; Skriptor, 1961, 41-55.

Zipf, G.K. The Psycho-Biology of Language. An Introduction to Dynamic Philology.-Boston: Houghton Mifflen, 1935.

Zsilka, T. Rekurentnost foném v poézii. - Metryka słowiańska.-Wrocław: Ossolineum , 1971.

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

Вектаев К. Б., Добин В. Г. Статистические приемы описания причин фонетических изменений в языке. - Известия Академии наук КазССР, серия общественная, 1969, № 1, с. 81-87.

Ведоногов Г.Г. Определение грамматических признаков "новых" слов с помощью словаря. - В кн.: Инженерная лингвистика. Л., 1971, с. 225-229.(Учен. зап. Ленинградского гос. ин-та им. А. И. Герцена. Том 458, ч. 2).

Вецозол И. Фонемная структура слова в современном английском и латышском языках. - Proceedings of the Sixth International Congress of Phonetic Sciences. Prague, 1967, p. 985-988.

Вецозол И. Количественное соотношение использования согласных и гласных фонем в слове в современном латышском и английском языках. - Балтийские языки и их взаимосвязи со славянскими, финно-угорскими и германскими языками. Тезисы докладов научной конференции, посвященной 100-летию со дня рождения акад. Я. Эндзелина. Рига, 1973, с. 83-86.

Вийтсо Т.-Р. Проблемы количества в эстонском языке. - Советское финно-угроведение XV, 1979, № 1, с. 1-17.

Гурдишева Е. А. Звуковой символизм и факторы, влияющие на него. - Учен зап. МГПИИЯ им. М. Тореза, 1973, т. 72.

Журавлев А. П. Фонетическое значение. - Л.: Изд-во ЛГУ,1974.

Зиндер Л. Р., Строева Т. В. Статистика гласных немецкого языка в диахронии. - В кн.: Исследования по немецкой филологии. Тула, 1972, с. 5-20.

Исенгельдина А. А. Некоторые вопросы фонологической статистики. - Вопросы языкознания, 1972, № 5, с. 119-129.

Курбаков К. И. Кодирование и поиск информации в автоматическом словаре. - М.: Советское радио, 1968.

Левицкий В.В. К проблеме звукосимволизма. – В кн.: Психологические и психолингвистические проблемы владения и овладения языком. М., 1969, с. 123-132.

Меликишвили И. Г. Восприятие фонем и некоторые универсальные закономерности частотных соотношений признаков фонем. – Материалы 3-го Всесоюзного симпозиума по психолингвистике. М., 1970, с. 162-163.

Милевский Т. Предпосылки типологического языкознания. – В кн.: Исследования по структурной типологии. – М.: Изд-во АН СССР, 1963, с. 3-31.

Никонов В. А. Интерпретация фонетических частот. – Ученые записки Ин-та славяноведения, 1963, т. 17, с. 259-270.

Перебейнос В. И. Использование статистических методов в типологических исследованиях (на материале славянских и германских языков). – В кн.: Проблемы языкознания. М. 1967.

Петрова Н. В. Кодовые характеристики письменного текста. – В кн.: Сатистика речи. / Отв. ред. Р. Г. Пиотровский. –Л.: Наука, 1968, с. 5-49.

Пиотровский Р. Г. Информационные измерения языка. – Л.:Наука, 1968.

Савченко И. Ф. Определение части речи словоформы украинского языка по конечной фонеме. – В кн.: Вопросы лингвостатистики и автоматизации лингвистических работ. М., 1970, с. 60-68. (Труды ЦНИИПИ. Серия 3/70. Выпуск 3).

Свецевичюс Б. И. К вопросу о частоте встречаемости фонем в литовской письменной речи. – Материалы коллоквиума лаборатории экспериментальной фонетики и психологии речи Ч. 2. Вильнюс, 1966, 19-22.

Статистико-комбинаторные моделирование языков. / Под ред. Н.Д. Андреева. – М.; Л.: Наука, 1965.

Статистичні параметри стилів. / Ред. В.С. Перебийніс. – Київ: Наукова думка, 1967.

Сухотин Б. В. Экспериментальное выделение классов букв с помощью электронной вычислительной машины. – В кн.: Проблемы структурной лингвистики. М., 1962, с. 198-206.

Тисенко Э. В. Алфавитно-графическая структура словаря. – Научный симпозиум "Семиотические проблемы языков науки, терминилогии и информатики". Ч. 1. – М.: Изд-во МГУ , 1971, с. 251-253.

Тулдава Ю. А. Квантитативное исследование структуры односложного слова в эстонском языке. – В кн.: Проблемы общей и прикладной лингвистики. Linguistica , X. Тарту, 1978, с. 115-135. (Учен. зап. ТГУ, вып. 453).

Фрумкина Р. М., Василевич А. П., Герганов Е. Н. Субъективные оценки частот элементов текста как прогнозирующий фактор. – В кн.: Вероятностное прогнозирование в речи. М. 1971, с. 70-93.

Хольм Х. А. Выделение первого морфологического типа в эстонском языке на основе статистико-комбинаторного моделирования в подъязыке радиоэлектроники. – В кн.: Статистико-комбинаторное моделирование языков. – М.; Л.: Наука, 1965а, с. 212-218.

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

ФОНЕТИКО-ГРАФЕМНЫЕ ИЗМЕРЕНИЯ ЭСТОНСКОЙ ЛЕКСИКИ

Юхан Тулдава

Р е з ю м е

В статье приводятся данные о частотности букв и звуков в эстонском тексте и словаре, а также данные о частотности начальных и конечных букв (звуков) в тексте. Констатируется наличие некоторых различий в распределении частот букв в текстах из разных подъязыков (см. табл. I, стр. 83). На основе суммарной выборки объемом в I63 340 букв (56% художеств. и 44% нехудожеств. текстов) пять самых частых букв (a, e, i, s, t) покрывают 49% и десять самых частых букв - 75% текста (табл. 4, колонка УI, стр. 86). Связь между частотой и рангом букв хорошо описывается логарифмической функцией: $p_r = a + b \lg r$ (p_r - частота, r - ранг, a и b - константы; для эстонского языка a = I3,7 и b = 9,6; см. рис. 2 на стр. 63). Устанавливается близость частотности букв в тексте и в соответствующем словаре (табл. 5 на стр. 87 и рис. 3 на стр. 64). Частотность звуков в тексте определяется на основе выборки из четырех подъязыков общим объемом в 23 560 звуков (соотв. 24 796 букв). Пять самых частых звуков (a, t/D, e, i, s) покрывают 53,6% текста (табл. 6, колонка III, стр. 88). Соотношение гласных и согласных звуков в тексте - 45,5:54,5 % (т.е. на I00 гласных приходится в среднем I20 согласных). Передние гласные составляют 50,3 %, задние (и средние) - 49,7% в тексте. Среди согласных 40 % смычных, 44,2 % щелевых и I5,8 % носовых звуков. Частотность начальных букв (табл. 8, стр. 90) и конечных букв (табл. 9, стр. 91) определяются на основе выборки из текстов четырех подъязыков. Устанавливается функциональная нагрузка букв в начальной и конечной позиции слова (табл. I0, стр. 92 и рис. 4, стр. 79). Статистические данные эстонского языка сопоставляются с соответствующими данными из других языков, в частности из родственных финского и венгерского языков.

PHONETIC-GRAPHEMIC MEASURES OF ESTONIAN LEXICS

Juhan Tuldava

S u m m a r y

The article presents statistical data on the frequency of letters (graphemes) and sounds in Estonian text and vocabulary as well as on the frequency of initial and final letters (sounds) of the words in a text. Some differences in the distribution of letters in the texts of various sublanguages are revealed (see Table 1, p. 83). On the basis of a summary sample of 163,340 letters (56 % fiction and 44 % non-fiction material) the five most frequent letters (a, e, i, s, t) cover 49 % and the ten most frequent letters - 75 % of the text (Table 4, Column VI, p. 86). The relation between the frequencies and the ranks of the letters can be expressed by the logarithmic function $p_r = a + b \lg r$ (p_r - frequency, r - rank, a and b - constants; for the Estonian text a = 13.7 and b = 9,6,see Fig. 2 on p. 63). Proximity between the frequencies of letters in a text and the corresponding vocabulary can be stated (Table 5 on p. 87 and Fig. 3 on p. 64). The frequency of the sounds in Estonian texts is determined on the basis of a summary sample from the texts of four sublanguages - the volume of the sample was 23,560 sounds (which corresponds to 24,796 letters). The five most frequent sounds (a, t/D, e, i, s) cover 53.6 % of the text (see Table 6, Column III, p. 88). The ratio between vowels and consonants is 45.5/54.5 % (i.e. 100 vowels to 120 consonants). The front vowels make up 50.3 %, the back (and medial) vowels - 49.7 % in the text. Of the consonants 40 % are occlusives, 44.2 % spirants, and 15.8 % nasals. The frequency of initial letters (Table 8 on p. 90) and final letters (Table 9 on p. 91) is determined on the basis of four sublanguages. The functional load of the letters in initial and final positions is calculated (Table 10 on p. 92 and Fig. 4 on p. 79). The statistical data on the Estonian language are compared with the corresponding data from other languages, in particular from the cognate Finnish and Hungarian languages.