Suomen kielen foneemien ja grafeemien frekvensseistä

OLLI JÄRVIKOSKI

Lauri Hakulinen julkaisi Suomen kielen rakenne ja kehitys -teoksensa ensimmäisen painoksen I osan (1941) alussa suomen äännerakennetta kuvastavat äänteiden suhteelliset frekvenssiluvut. Lukujen mukaan suomen kielen yleisin äänne on *i* ja sitä seuraavat kymmenen äännettä ovat *t*, *a*, *e*, *s*, *n*, *ä*, *l*, *k*, *o* ja *u* esitetyssä järjestyksessä. Ensimmäisen kerran Hakulinen oli julkaissut nämä luvut Virittäjässä 1938. Hakulinen ei selvitä aineistonsa laatua eikä laajuutta. Se ei kuitenkaan ole varmaan voinut olla kovin laaja. Toisaalta äänteet ovat niin pieniä kielellisiä yksiköitä, että jo suhteellisen suppeastakin korpuksesta voidaan saada oikean suuntaisia tuloksia.

Kaikitenkin nämä Hakulisen tiedot pysyivät 30 vuotta sinä aineistona, johon suomen kielen äänteiden — tai foneemien — suhteellisista taajuuksista keskusteltaessa luotettiin. Hakulinen julkaisi samat tulokset vielä SKRK:n kolmannessa painoksessa vuonna 1968. Vasta 1970-luku horjutti Hakulisen antamaa kuvaa. ATK-menetelmien kehitys ryöpsäytti esiin koko joukon tutkimuksia, joissa suomen äänteiden ja kirjainten useuksia selvitettiin. Kun seitsenkymmenluvun alkupuolella luin ensimmäisiä tutkimuksia, joissa Hakulisen äänneluettelo osittain kyseenalaistettiin, alkoi tämä asia kiinnostaa. Pari kolme vuotta sitten tarjoutui tilaisuus Lauseopin arkiston ATK-järjestelmää kehiteltäessä saada tietoja arkiston murreaineiston foneettisten merkkien ja toisaalta — tuolloin varsin suppean — kirjakielisen materiaalin grafeemien frekvensseistä. Silloin ryhdyin valmistelemaan tätä kirjoitusta. Perustulostukset omia tarkempia laskelmiani varten hoiti Turun yliopiston DEC 20 -tietokoneella Lauseopin arkiston ATK-asiantuntija ja ohjelmiston rakentaja maisteri Pekka Porri.

1. 1970-luvun tutkimukset: *a*:sta *i*:n haastaja

Aloitan katsauksella seitsenkymmenluvun tutkimuksiin. Eri tutkijoiden termistö ja laskentaperusteet ovat vaihdelleet jonkin verran, mutta tulokset ovat silti hyvin vertailtavissa. Myös heidän käyttämänsä korpukset ovat

olleet erilaisia, mikä antaa hyvät mahdollisuudet ryhtyä luomaan kokonaiskuvaa.

Hakulisen tarkoituksena oli siis antaa tietoja suomen ääntei s t ä. Jaakko Pesonen (1971) laski k i r j a i m i a sanomalehtiteksteistä. Hänen otoksensa oli kirjainfrekvenssejä laskettaessa lähes 175 000 kirjainmerkin suuruinen. (Pesonen 1971: 5—6.) Pesosen käytettävissä oli jo Valde Mikkosen vuodelta 1969 peräisin oleva käsikirjoitus, joka ilmestyi painettuna vasta vuotta Pesosen julkaisun jälkeen (1972). Mikkonen ilmoitti laskevansa m e r k k e j ä, joihin hän laski myös sananvälin. Hänen aineistonsa koostui kolmesta yhtä suuresta otoksesta, jotka olivat peräisin aapisesta, Kalevalasta ja Väinö Linnalta; otoksen koko laajuus oli noin 24 000 merkkiä. (Mikkonen 1972: 21—26.) Mikkosen ja Pesosen tulokset ovat hyvin samankaltaiset. Molemmat päätyivät kirjaintilastollaan kumoamaan Hakulisen äännetilaston: molempien tutkimusten mukaan kuusi ensimmäistä kirjainta ovat järjestyksessä *a*, *i*, *t*, *n*, *e*, *s*.

Jo samana vuonna 1972 julkaisi Vilho Setälä perusteellisen tutkimuksen »suomen kielen dynamiikasta». Setälän aineistona oli Uusi testamentti, ja otos oli 766 000 kirjaimen suuruinen. (Setälä 1972: 8.) Setälän kuuden ensimmäisen k i r j a i m e n järjestys on yhtäpitävä Mikkosen ja Pesosen kanssa paitsi yhdeltä kohdin: *n*:n frekvenssi on ohittanut *t*:n. Setälä asettaa (mts. 12) Hakulisen tulokset erityisesti *i*:n yleisimmyyden osalta perusteellisesti kyseenalaisiksi. Setälä julkaisee myös kirjaintilaston pohjalta muunnetun äännetilaston, joka hänen mukaansa samalla on foneemitilasto (mts. 19, perusteluja ks. s. 8 ja 20). Siinä hän esittää erikseen lyhyiden ja pitkien ä ä n t e i d e n tilastot sekä pitkien ja lyhyiden yhteen lasketun taulukon, joissa hän tyytyy esittämään 16 yleisintä äännettä. Niistä jälleen kuusi ensimmäistä (nyt eri järjestyksessä) ovat *a*, *n*, *i*, *e*, *t*, *s* (sekä lyhyiden taulukossa, josta pitkät on poistettu, että yhteistaulukossa, jossa pitkiäkin edustaa vain yksi merkki). (Mainittakoon, että pitkien kuusi ensimmäistä ovat *t*, *l*, *a*, *i*, *e*, *ä*.)

Matti Pääkkösen aineisto vuodelta 1973 on tähän mennessä laajin. Pääkkönen laskee g r a f e e m e j a, mutta hänen tuloksensa ovat täysin vertailukelpoiset kaikkien edellä mainittujen kirjaintilastojen kanssa. Grafeemeja kertyy Pääkkösen otokseen kirjakielestä lähes 2 500 000 ja yleispuhekielestä melkein 640 000; yhteismääräksi saadaan huikeat yli 3 130 000. (Pääkkönen 1973: 318—319.) Pääkkönen siis laskee grafeemeja myös puhekielestä; joka tapauksessa hänen tilastonsa käyvät kuuden ensimmäisen grafeemin osalta molemmissa aineistoissa täysin yksiin Mikkosen ja Pesosen kanssa: *a*, *i*, *t*, *n*, *e*, *s* (*n* siis vasta *t*:n jäljessä).

Kaisa Häkkisen (1977) aineistona oli suomalainen kansansatusovitelma, jonka laajuus on noin 140 tekstisivua. Häkkinen valitsi korpuksensa pyrkien

välttämään tekstejä, jotka sisältäisivät nuoria lainasanoja. Häkkinen haluaa laskea nimenomaan f o n e e m e j a , vaikka hänen aineistonsa lienee lähinnä kirjakieltä; foneemien laskemiseen hän uskoo Hakulisenkin tähdänneen. Häkkisen tulos (1977: 57—58, 63) on kuuden ensimmäisen foneemin osalta täsmälleen sama kuin Setälän kirjainluettelo: *a*, *i*, *n*, *t*, *e*, *s*.

Tällaisen todistusaineiston edessä Hakulinen laski aseensa. Hän puhuu SKRK:n 4. painoksessa (1979) äänteiden sijasta foneemien taajuusluvuista ja esittää »uusimpien tilastojen mukaan» suomen yleisimmäksi foneemiksi *a*-vokaalin ja toiseksi yleisimmäksi *i*-vokaalin. Samalla hän esittää rinnakkain prosenttilukuineen Setälän kirjaintilaston ja Pääkkösen grafeemitilaston sekä vertailee näitä keskenään (mts. 18—21). Tämä jäikin viimeiseksi sanaksi 1980-luvulle tultaessa.

2. Foneemin ja grafeemin käsitteestä

Edellä on laskettu milloin äänteitä tai foneemeja, milloin kirjaimia tai grafeemeja. Tuloksia on myös yleistetty näiden rajojen yli. Kirjaimen ja grafeemin käsitteiden voidaankin katsoa lähes samastuvan, jollei haluta lähteä saman kaltaiseen »grafemaattiseen» tulkintaan, joka Setälällä on fonologisena: hän erottaa alustavasti pitkät ja lyhyet äänteet (ja foneemit) toisistaan (sekä vokaalit että konsonantit). Tästä seuraa, että vaikka Häkkisen foneemiluettelo on alkupäästään identtinen Setälän kirjainluettelon kanssa, se ei samalla tavoin yhdy Setälän foneemi- ja äänneluetteloon! Setälän foneemitulkinta on toki mahdollinen. Se on kuitenkin vain yksi kolmesta vaihtoehtoisesta tulkintamahdollisuudesta eikä niistä paras, kuten Fred Karlsson (1969, erit. s. 353—355) on osoittanut. Häkkinen (1977) ja Hakulinen (1979) edustavatkin tätä perustellumpaa fonologista linjaa, jonka mukaan pitkien vokaalien ja konsonanttien katsotaan sisältävän kaksi foneemista yksikköä diftongien tapaan. Tulen edustamaan samaa tulkintaa omissa foneemitilastoissani, jotka julkaisen tuonnempana. Vertailun kannalta muistettakoon lisäksi, että vanhoissa murteista lasketuissa äännefrekvensseissä (esim. Ruoppila 1936 ja Lepistö 1938) on puolikarkean tarkekirjoituksen periaatteita noudattaen pitkät vokaalit laskettu kauttaaltaan yhdeksi äänteeksi (mutta sen sijaan geminaattakonsonantit kahdeksi).

Häkkinen (mts. 62—63) polemikoi lisäksi »polysysteemisen» foneemikäsityksen puolesta. Hänen artikkelissaan tämä ongelma rajoittuu kysymykseen *ŋ*:stä eli siitä, voiko *ŋ* samaan aikaan olla foneemi ja */n/*:n allofoni. Häkkinen lähtee laskelmissaan siitä, että vain tavunalkuinen *ŋ* on fonologisesti */ŋ/*. Hän laskee silti */ŋ/*-foneemille kaksi taajuuslukua, sekä tavunalkuisen *ŋ*:n että kaikkien *ŋ*:n esiintymien pohjalta. Vaikka tunnen jonkinlaista houkutusta polysysteemiseen tulkintaan, pysyn silti Häkkisen »fysikaaliseksi»

nimittämässä foneemitulkinnassa. Foneettisesti litteroitu murreaineistoni antaisi kyllä mahdollisuudet polysysteemiseen tulkintaan ja tähän perustuvaan laskentaan. Esim. /*n*/:n allofoneja tällöin kuitenkin murreaineistosta löytyisi paljon runsaammin kuin pelkästään osasta *ŋ*-tapauksia; niitä olisivat esim. *m*, *l* ja *v* seuraavissa assimilaatiotapauksissa: *mennäämpäs*, *otinkim pois*, *sillal luo*, *puhutaav vaa*. Katson siis /*m*/:n, /*l*/:n ja /*v*/:n näissä tapauksissa edustavan niitä foneemeja, joiksi ne on jo muussa yhteydessä suomen kielessä ja asianomaisessa murteessa todettu.

On kuitenkin ilmeistä, että valitaanpa laskelmien pohjaksi polysysteeminen tai monosysteeminen foneemikäsitys, tulokset eivät mitenkään olennaisesti muutu ainakaan foneemiluettelon alkupäästä. Tätä arviota tukevat myös Setälän tarkistuslaskelmat (mts. 20). Ongelmallisempia ovat ainakin periaatteessa yleistykset aineistosta toiseen grafeemin- ja foneeminrajan yli. Voidaan kysyä, onko teoreettisesti edes suotavaa laskea puhekielestäkin grafeemeja Pääkkösen tapaan. (Pääkkösen käytännön tulokset sekä kirjaettä puhekielestä ovat kylläkin hyvin samanlaiset, ja hän voi näin antaa grafeemeina tietoja myös puhekielen foneemeista.) Setälä ja Häkkinen taas ovat — tietäen aineistonsa kohtuullisesti kirjallisen luonteen — lähteneet grafeemien avulla saaduista tuloksista yleistyksiin äänne- tai foneemitaajuuksiksi. Ja lopuksi myös Hakulinen on nämä tulokset hyväksynyt.

Käytännössä tällaisin yleistyksin saaduissa tuloksissa ei tietenkään ole vikaa — jos ne vain ovat oikeita. Ongelmaksi saattaa kuitenkin nousta kysymys siitä, mitä lopulta ovat »suomen kielen» foneemitaajuudet. Mikä on se suomen kieli, josta nämä taajuudet olisi mitattava? Löytyykö mahdollisesti eri kielimuotoja, joissa nämä useudet ovat erilaiset, ja jos, mikä kielimuoto tai millainen otos eri kielimuodoista olisi valittava edustamaan »suomen kieltä»? Parempi tapa ehkä olisi pyrkiä lopputulokseen asteittain, ensin eri kielimuotoja tarkastellen.

3. Otoksen luotettavuus ja yleistettävyys

Tällaisten tutkimusten yksi peruskysymys liittyy siis aineistoon — se kysymys, onko valittu korpus luotettava, reliaabeli, eli edustaako korpus oikealla tavalla sitä perusjoukkoa, jota sen tulokset on määrä yleistää edustamaan.

Lauseopin arkiston lopullinen murreaineisto edustaa kaikkia suomen murrealueita niin, että jokaiselta alueelta joka neljännestä pitäjästä on yksi noin 6 000—7 000 saneen murrenäyte (joiltakin alueilta on tiheämpikin seula). Aineiston pohjana ovat Turun ja Helsingin yliopistoissa tehdyt vanhat murreäänitteet, joista otos on valikoitu pitäen erityisesti silmällä murteen ja puhunnan laatua. Äänitteissä vallitseva puhetilanne on luonteeltaan haastatte-

lu, mutta LA:n aineisto on pyritty mahdollisuuksien mukaan valitsemaan niin, että tämä puoli ei korostuisi, vaan murteenpuhujan autenttinen vapaa puhunta ja hänen itsensä sommittelema vapaa kerronta pääsisi oikeuksiinsa. Aineiston luonteesta kyllä seuraa, että se tekstuaalisesti pääfunktioltaan edustaa kertovaa ilmaisua, vaikka juuri tähän ei mitenkään ole pyritty, pikemminkin päinvastoin.

Aineisto on kerätty ensi sijassa syntaktisia tarkoituksia varten. Alusta asti on kuitenkin tiedetty, että sillä on merkitystä myös muunlaisten tutkimustehtävien kannalta, ja arkiston lopullisen ATK-järjestelmän nyt valmistuessa tämä on yhä ilmeisempää. On toisaalta selvää, että sovellettaessa tätä materiaalia muihin kuin puhtaasti syntaktisiin tehtäviin siinä saattaa ilmetä puutteita. Aineiston mahdolliset rajoitukset uuden tutkimustehtävän kannalta on siis aina erikseen harkittava.

Lähden joka tapauksessa siitä, että tutkimukseni taustalla oleva perusjoukko on murreaineistoni osalta 1950—70-luvulla tallennettujen suomen murteiden foneemikanta. Eräänlainen toisen asteen perusjoukko oman otokseni kannalta on LA:n koko murreaineisto; tähän liittyy kysymys otokseni validiudesta. Olen poiminut otokseni eri murteita edustavista 40 murretekstistä. (Oikeastaan tekstejä on tarkasti 42; ks. artikkelin lopussa olevaa liitettä.) Nämä noin 40 tekstiä olen valinnut tasaisesti eri murrealueilta LA:n luokituksen mukaan siten, että kutakin aluetta edustaa kaksi näytettä.[1] Murrealueiden sisällä pitäjittäinen murrenäytteen valinta ei voinut olla murremaantieteellisesti aivan ihanteellinen, mikä johtuu siitä, että tulostukset tehtiin välivaiheen tiedostoista, joissa koko aineisto ei vielä ollut mukana. Olen kuitenkin pyrkinyt siihen, että näytteet edustavat kvaliteettinsa ohella myös kvantitatiivisesti mahdollisimman tasaisesti eri murrealueita, jotta luotettava, kaikki murrealueet kattava kokonaiskuva mahdollistuisi. Yhdessä nämä nelisenkymmentä murretekstiä ovat vajaa kolmannes LA:n kaikkien murretekstien määrästä. Rajasin lopulliseksi otokseksi kyseisten murretekstien alkuosan, 10 000 graafia[2] kustakin (kuitenkin siten, että sanetta ei tarvitse katkaista). Tämä merkitsee ehkä hiukan vajaata kolmasosaa kyseisen tekstimäärän koko laajuudesta (yksityisten murretekstien graafi-

---

[1] Murreaineistoni käsittää 10 000 graafin otokset pitäjää kohti eri murrealueilta. Näistä olen aputulostusten pohjalta laskenut esiin foneemit (tämä koskee /*ŋ*/-tapauksia ja itämurteiden palataalistuneita konsonanttifoneemeja). Ks. artikkelin lopussa olevaa liitettä.

[2] *Graafi* tarkoittaa murreaineistosta puhuessani niitä LA:n karkean tarkekirjoituksen äännemerkkejä, jotka usein jonkin verran foneettisista merkeistä muunnetussa asussa on tallennettu tietokoneella luettaviksi. Graafeja ovat esim. A = *a*, N, = *ń*, TX = *ϑ*, DX = *δ* ja Q = ʔ tai '. Digrafi NG (= *ŋŋ*) sisältää kaksi graafia (ja foneemia). Mainittakoon, että dentaalispirantteja merkitseviä graafeja ei otoksessani ollut lainkaan. Muussa yhteydessä *graafia* voidaan käyttää *kirjaimen* synonyyminä.

määrä näkyy vaihtelevan vähän alle 30 000:sta jonkin verran sen yli). Lopullinen otokseni edustaa siis 1/9—1/10:aa LA:n tasaisesti eri murteita edustavasta aineistosta (LA:n koko materiaali on tätä hiukan laajempi). Koska foneemit muodostavat umpiluokkia ja niitä pieninä yksikköinä mahtuu tällaiseen otokseen suuria määriä, otokseni voidaan katsoa edustavan hyvin LA:n koko aineistoa (vrt. Kohonen—Salmela 1978: 7—11). Myös osajoukkojen myöhemmin taulukossa 3 osoittama perussuuntaus tukee tätä käsitystä. En puutu tässä tarkemmin asian tilastotieteelliseen puoleen. Koko murreotokseni käsittää siis yhteensä yli 400 000 graafia.

Murreaineistostani erillinen kirjakielen aineistoni on paljon suppeampi: se on edellisestä vain noin kymmenesosan eli 40 000 graafin suuruinen. Oletin kuitenkin, että jo näin laaja otos voi antaa luotettavan kuvan kirjakielen

TAULUKKO 1. Eri tutkimusten 20 yleisintä äännettä tai foneemia taikka kirjainta tai grafeemia useusjärjestyksessä.

Ruoppila 1936 (10 000):																				
Lemin murre	i	e	a	t	s	o	k	l	n	u	ä	m	p	h	v	r	y	j	ö	ŋ
Lepistö 1938[1] (14 000):																				
Vampulan murre	i	t	a	l	k	s	n	e	m	o	ä	u	p	r	j	v	h	y	n	ö
Hakulinen 1938—68 (?):																				
?	i	t	a	e	s	n	ä	l	k	o	u	—	—	—	—	—	—	—	—	—
Pesonen 1971 (175 000):																				
sanomalehdet	a	i	t	n	e	s	l	o	k	u	ä	m	v	r	j	h	p	y	d	ö
Mikkonen 1972 (3 x 8 000 = 24 000):																				
aapinen	a	i	t	n	e	s	k	u	l	o	ä	p	m	r	h	v	j	y	ö	d
Kalevala	a	i	e	n	t	l	s	k	u	ä	o	m	v	h	r	p	y	j	ö	g
Linna: Pohjantähti	a	i	t	n	e	s	k	l	ä	u	o	m	h	p	r	v	j	y	d	ö
koko aineisto	a	i	t	n	e	s	l	k	u	o	ä	m	h	v	p	r	j	y	ö	d
Setälä 1972 (766 000):																				
UT: kirjaimet	a	i	n	t	e	s	ä	l	k	o	u	m	h	j	v	p	r	y	d	ö
lyhyet äänteet	a	n	i	e	t	s	ä	o	k	u	l	m	(p	r	y	ö)	—	—	—	—
lyhyet + pitkät äänteet	a	n	i	e	t	s	ä	k	o	l	u	m	(p	r	y	ö)	—	—	—	—
Pääkkönen 1973 (2 500 000 + 640 000 = 3 130 000):																				
kirjakieli	a	i	t	n	e	s	l	k	o	u	ä	m	v	r	j	h	y	p	d	ö
yleispuhekieli	a	i	t	n	e	s	l	o	k	ä	u	m	v	r	j	h	y	p	d	ö
koko aineisto	a	i	t	n	e	s	l	k	o	u	ä	m	v	r	j	h	y	p	d	ö
Häkkinen 1977 (140 sivua):																				
kansansatu	a	i	n	t	e	s	k	l	ä	u	o	m	p	v	r	j	h	y	ö	d
Järvikoski 1984 (40 000 ja 400 000):																				
kirjakieli: grafeemit	a	i	t	n	e	s	l	u	k	o	ä	m	r	v	j	h	p	y	d	ö
murteet: foneemit	i	t	a	e	s	n	l	k	o	ä	m	u	j	v	h	p	r	y	ŋ	ö

[1] Lepistön tutkimuksessa *i*:n ja *a*:n frekvenssi (varsinkin kun hänen erikseen laskemiensa *ī*:n ja *ā*:n vaikutus eliminoidaan) käy aivan yksiin (ja *a* ohittaa *t*:n).

grafeemien taajuusjärjestyksestä, jos otos edustaa kyseistä kielimuotoa riittävän monipuolisesti. Otos sisältää lehtikieltä ja historian alaan kuuluvaa tekstiä. (Olisin kyllä ottanut kirjakielestäkin hiukan laajemman otoksen, jos tiedostoja olisi ollut käytettävissä.)

4. Suomen äännerakenteen perusta: kuusi vokaalia — kuusi konsonanttia

Olen koonnut taulukkoon 1 kaikkien tähän mennessä mainitsemieni tutkimusten äänne-, kirjain-, foneemi- ja grafeemiluettelot esitettyinä yleisimmästä 20:nneksi yleisimpään. Ylimpänä ovat Ruoppilan ja Lepistön vanhat yhden paikallismurteen äännetutkimukset, alimpana omat tulokseni. Asetan rinnalle taulukon 2, jossa ovat eri murrealueiden foneemien yleisyysluettelot erikseen.

Taulukosta 1 havaitaan konkreettisesti kuuden ensimmäisen foneemin ja grafeemin ryhmä: kolme vokaalia *a*, *e* ja *i* sekä kolme konsonanttia *n*, *s* ja *t*. Vain vanhimmissa murretuloksissa tämä ryhmä hajoaa: Ruoppilalla *n* jää 9:nneksi (vaikka siihen sisältyy *ń*:nkin osuus) ja Lepistöllä *k* ja *s* ohittavat *n*:n ja *e*:n (ja Mikkosen Kalevala-aineistossa *l* nousee *s*:n ohi). Taulukosta 2 näkyy, että aineistossa kKM/b *l* niin ikään on korkealla, jopa 5:ntenä, otoksessa eKM/b 6:ntena. Havaitaan myös, että murteittain tämä »kuuden ryhmä» tosin kuusi kertaa hajoaa, mutta vain kolme kertaa (LM:ssa) jokin sen jäsen (*n*) jää 7. sijaa alemmaksi. Tätä kolmen vokaalin ja kolmen kon-

TAULUKKO 2. Foneemien taajuusjärjestys 20 murreryhmässä. Laskelmissa oli mukana laryngaaliklusiili-puristussupistuma fonologisen asemansa epämääräisyydestä huolimatta.

LM/a	i	t	a	s	e	k	o	n	l	m	ä	u	j	p	v	h	r	y	ŋ	ö
LM/b	i	t	s	a	e	k	l	n	o	ä	u	m	j	p	r	v	h	y	ŋ	ö
LM/c	i	t	s	a	k	e	o	l	u	n	m	ä	v	j	p	r	h	y	ŋ	ö
LsM	i	a	t	e	s	n	l	ä	k	o	u	m	j	v	h	r	p	y	d	ŋ
pHM/a	i	t	a	e	n	s	ä	k	l	o	m	u	j	p	r	v	h	y	ŋ	ö
pHM/b	i	t	a	s	e	n	ä	l	k	o	m	u	p	j	v	r	h	y	ŋ	ö
eHM/a	i	a	t	n	s	e	l	ä	o	k	u	m	j	v	h	r	p	y	ŋ	ö
eHM/b	i	a	t	n	e	s	k	l	ä	o	m	u	j	v	h	p	y	r	ŋ	ö
ePM	i	a	t	n	s	e	l	ä	o	k	u	m	j	h	r	v	p	y	ŋ	ö
kPM	i	a	t	e	n	s	o	l	ä	k	u	m	j	h	v	p	r	y	ŋ	ö
pPM	i	a	t	e	n	s	l	o	k	ä	u	m	j	v	p	h	y	r	ŋ	’
PpM	i	t	a	e	n	s	l	o	k	ä	u	m	h	j	v	p	y	r	ŋ	ö
leSM/a	i	e	t	a	s	n	ä	o	l	k	m	u	j	v	p	h	r	y	ŋ	ö
leSM/b	i	a	e	t	n	s	o	ä	k	l	u	m	h	v	j	p	r	y	ŋ	ö
ipSM/a	i	e	a	t	s	n	ä	o	l	k	m	u	v	j	h	p	r	y	ö	ŋ
ipSM/b	i	e	t	a	s	n	o	ä	l	k	m	u	j	v	p	h	r	y	ŋ	ö
eKM/a	i	t	a	e	s	n	o	l	k	ä	u	m	h	j	v	p	y	r	ŋ	ö
eKM/b	i	t	a	e	s	l	n	o	k	ä	m	u	j	p	h	v	r	y	ŋ	ö
kKM/a	i	t	a	e	o	s	n	k	l	ä	u	m	j	h	v	y	p	r	ŋ	ö
kKM/b	i	t	e	a	l	s	n	o	ä	k	m	u	h	v	j	p	y	r	ö	ŋ

sonantin ryhmää voidaan siis varsin vahvasti pitää suomen kielen foneemien ja grafeemien eräänlaisena perusryhmänä.

Edellistä vielä huomattavasti selvempi ryhmä syntyy 12 ensimmäisestä foneemista ja grafeemista. Tähän ryhmään kuuluvat edellisten lisäksi vokaalit *o*, *u* ja *ä* sekä konsonantit *k*, *l* ja *m*. Tämä ryhmä on täysin yhtenäinen kaikissa tähänastisissa tutkimuksissa paitsi yhtä pientä osa-aineistoa, Mikkosen aapistilastoa, jonka edustavuutta koko kielen kannalta voi hiukan epäillä, sekä samoin jokikisellä murrealueella (ks. taulukkojen 1 ja 2 katkoviivaa!). Vilkaistakoon jo nyt myös taulukkoa 3. Siitä näkyy, että yleensä 12. foneemin jälkeen taajuus prosenttisesti vähenee selvästi (noin 1—1 1/2, jopa yli 2 %); tämä ei koske pohjalaismurteita. Voimme nyt lopullisesti tietää, mitkä ovat nykysuomen 12 yleisintä foneemia ja grafeemia.

5. Murteiden *i* ja kirjakielen *a* vastakkain

Kolmas taulukoista paljastuva tärkeä seikka koskee *a*:n ja *i*:n frekvenssin suhdetta. Taulukosta 1 näkyy, että ne luettelot, joissa *a* on ensimmäisenä,

TAULUKKO 3. Foneemien useusjärjestykset ja useusprosentit kymmenessä eri murreryhmässä. Prosentit ilmoitetaan yleensä kymmenesosan tarkkuudella, mutta milloin on jostakin syystä tarpeen, sadasosan. Tuhannesosaprosentteja ei yleensä oteta huomioon. Taulukon yhtenäisyyden vuoksi on itämurteiden liudentuneet konsonanttifoneemit sisällytetty liudentumattomiin, mutta näiden rinnalla sulkeissa mainitaan liudentuneiden foneemien prosenttiluvut, joten kyseisten murteiden foneemien todellinen frekvenssi on taulukosta laskettavissa.

	LM			LsM			pHM			eHM	
1	i	12,4	1.	i	13,7	1.	i	12,7	1.	i	12,8
	t	10,0		a	10,3		t	10,1		a	9,8
	a	9,00		t	9,7		a	9,6		t	9,4
	s	8,99		s	7,7		e	8,2		n	8,3
	e	7,2		e	7,2		s	7,9		e	7,8
6.	k	6,8	6.	n	7,1	6.	n	7,4	6.	s	7,7
	l	6,1		k	6,2		ä	6,8		l	6,0
	l	6,0		l	6,1		l	5,9		k	5,8
	n	5,5		ä	5,9		k	5,7		ä	5,7
	m	4,93		o	5,7		o	5,1		o	5,4
	ä	4,87		m	4,1		m	4,7		m	4,4
12.	u	4,7	12.	u	4,0	12.	u	3,7	12.	u	4,2
	j	2,8		j	2,5		p	2,18		j	2,4
	p	2,1		v	2,2		j	2,16		v	2,2
	v	2,0		p	1,8		r	1,83		h	2,0
	r	1,9		r	1,71		v	1,79		p	1,8
	h	1,6		h	1,66		h	1,6		r	1,7
	y	1,5		y	1,3		y	1,4		y	1,6
	ŋ	1,1		ŋ	0,8		ŋ	0,9		ŋ	1,0
20.	ö	0,2	20.	ö	0,15	20.	ö	0,2	20.	ö	0,2
	f	0,09		d	0,08		d	0,02			
	d	0,01		f	0,05						

	ekPM		pP&PpM		leSM	
1.	i	13,2	i	12,4	i	11,3
	a	9,8	a	10,0	e	9,7
	t	9,3	t	9,2	t	9,31 (t' 0,03)
	e	8,03	e	8,4	a	9,1
	n	7,96	n	7,8	s	7,84 (ś 0,03)
6.	s	7,3	6. s	7,7	6. n	7,29 (ń 0,05)
	l	6,5	l	6,6	ä	6,272 [!]
	o	6,1	o	6,0	o	6,267 [!]
	ä	5,9	k	5,7	l	5,9 (l' 0,9)
	k	5,0	ä	5,6	k	5,7
	u	3,9	u	4,1	u	4,7
12.	m	3,3	12. m	3,4	12. m	4,3
	j	3,1	j	2,8	j	2,5
	h	2,7	h	2,4	v	2,2
	v	2,0	v	2,0	h	2,0
	p	1,8	p	1,9	p	1,8
	r	1,3	y	1,6	r	1,4
	y	1,3	r	1,2	y	1,3
	ŋ	0,8	ŋ	0,8	ŋ	0,7
20.	ö	0,2	20. ö	0,3	20. ö	0,3
	ʔ	0,05 [kPM]	ʔ	0,17 [pP]	ʔ	0,05
	f	0,02 [ePM]				

	ipSM		eKM		kKM	
1.	i	11,4	1. i	13,9	1. i	14,1
	e	9,8	t	10,1 (t' 0,1)	t	10,1 (t' 0,1)
	a	9,2	a	9,5	e	8,90
	t	8,99 (t' 0,04)	e	9,2	a	8,86
	s	8,0 (ś 0,2)	s	8,53 (ś 0,01)	s	6,82 (ś 0,03)
6.	n	7,8 (ń 0,1)	6. n	6,19 (ń 0,06)	6. n	6,64 (ń 0,01)
	o	6,2	l	5,94 (l' 0,4)	o	6,58
	ä	6,1	o	5,93	l	6,49 (l' 0,6)
	l	5,8 (l' 0,6)	k	5,6	k	5,8
	k	5,4	ä	5,3	ä	5,2
	m	4,2	u	4,0	m	3,9
12.	u	3,9	12. m	3,9	12. u	3,8
	j	2,7	j	2,3	h	2,5
	v	2,5	h	2,1	v	2,4
	h	2,20	p	1,833 [!]	j	2,250
	p	2,18	v	1,830 [!]	y	1,83
	r	1,59 (ŕ 0,02)	y	1,54	p	1,77
	y	1,4	r	1,46	r	1,4
	ŋ	0,7	ŋ	0,45	ŋ	0,63
20.	ö	0,5	20. ö	0,38	20. ö	0,58
	ʔ	0,28				

pitänee yleensä tulkita kirjakieltä edustaviksi. Selvä poikkeus tästä on Pääkkösen yleispuhekielen aineisto ja niin ikään Mikkosen Kalevala-aineisto. *a*:n kiistaton asema kirjakielessä näkyy myös omista tuloksistani, jotka ovat taulukossa alimpana. Erikoisasemaan joutuu Hakulisen aineisto, jonka laatu ei ole täsmällisesti tiedossamme. Hakulisen äänneluettelo on kuuden ensimmäisen osalta identtinen murteiden foneemiluetteloni kanssa ja poikkeaa selvästi kaikista kirjakieleen perustuvista luetteloista (myös omastani). Jo

vuoden 1938 artikkelinsa Hakulinen aloittaa luettelemalla suomen yleiskielen äänteet, joiden frekvenssejä hän näin ollen artikkelissaan tarkastelee. Uudemmat tutkimustulokset panevat kuitenkin kysymään, mitä Hakulisen otos täsmällisesti sisältää: sisältyykö hänen aineistoonsa sekä puhe- että kirjakieltä ja missä suhteessa?

Taulukosta 2 taas huomataan, että *i*:n asema kaikkien murteiden yleisimpänä foneemina on aivan vankkumaton. Mikään muu foneemi ei ole kaikkien murteiden listalla pysyvästi samalla sijalla. Taulukko 3 kertoo lisäksi, että myös prosentteina mitaten *i*:n ero seuraavaan foneemiin (*t*:hen, *a*:han tai *e*:hen) on hyvin selvä.

Näyttää siis syntyvän selvä jako: *i* on (murteiden) yleisin foneemi ja *a* (kirjakielen) yleisin grafeemi. Näiden tulosten perusteella 1970-luvun lopussa omaksuttu kanta *a*:n ehdottomasta valta-asemasta suomen foneemien joukossa ei enää näytäkään selvältä.

Mikä selittäisi *i*:n ylivertaisuuden murteissa? Vaikuttavin syy tähän lienee loppuheitto. Kirjakielessä *a* on (*n*:n jälkeen) toiseksi yleisin sananloppuinen grafeemi, ja sen frekvenssi on tulosteni mukaan peräti 25,2 %. Murteissa loppu-*a*:n osuus on vain 16,8 %, mutta loppu-*i*:n suhteellinen frekvenssi nousee vastaavasti kirjakielen 10,3:sta murteiden 13,9 prosenttiin. Myös muita selityksiä tähän lienee löydettävissä.

TAULUKKO 4. Foneemien ja grafeemien taajuusprosentit.

	Foneemit (murreaineisto):					Grafeemit (kirjakieli):				
1.	i	12,8		j	2,6	1.	a	12,1		r	2,5
	t	9,6		v	2,12		i	10,9		v	2,4
	a	9,5		h	2,05		t	9,6		j	2,1
	e	8,4		p	1,9		n	8,9		h	2,0
	s	7,9		r	1,6		e	7,83		p	1,9
6.	n	7,8		y	1,5	6.	s	7,79		y	1,6
	l	6,1		ŋ	0,8		l	5,8		d	0,9
	o	5,9	20.	ö	0,3		u	5,3	20.	ö	0,5
	k	5,8		(ʔ | 0,05)		k	5,22		g	0,2
	ä	5,7		(f | 0,02)		o	5,17			
	m	4,149		(d | 0,01)		ä	4,2			
12.	u	4,124				12.	m	3,0			

Taulukossa 4 julkaisen foneemien ja grafeemien taajuusprosentit koko aineistostani. Kun verrataan foneemien prosenttilukuja Hakulisen (esim. 1968: 16) lukuihin, huomataan kärjen osalta seuraavat suhteet: *i* 12,8/12; *t* 9,6/11,5; *a* 9,5/10,4; *e* 8,4/9,4; *s* 7,9/8,5; *n* 7,8/8,4. Minulla siis *i*:n asema on korostuneempi muihin nähden kuin Hakulisella; selityksenä lienee toisaalta se, että Hakulisen aineisto sisältänee kirjakieltä (jolloin *a*:n ja *n*:nkin asema

vahvistuu), toisaalta oman aineistoni mahdolliset rajoitukset.[3] Katson kuitenkin, että taulukko 4 kertoo sen olennaisen eron, joka vallitsee murretta käyttävän vapaan puhunnan foneemitaajuuksien ja kirjakielen grafeemitaajuuksien välillä.

TAULUKKO 5. Vokaalien prosenttiosuudet kymmenessä murreryhmässä sekä kirjakielessä.

	LM:	LsM:	pHM:	eHM:	ekPM:	pP&PpM:
a	19,6	21,3	20,0	20,8	20,2	20,7
e	15,7	15,0	17,1	16,4	16,5	17,4
i	27,1	28,3	26,7	27,0	27,2	25,6
o	13,1	11,9	10,7	11,3	12,7	12,5
u	10,3	8,2	7,8	8,8	8,1	8,4
y	3,2	2,8	3,0	3,3	2,6	3,4
ä	10,6	12,3	14,2	12,1	12,1	11,5
ö	0,4	0,3	0,5	0,4	0,7	0,6

	leSM:	ipSM:	eKM:	kKM:	Kirjakieli:
a	18,7	19,0	19,1	17,8	25,3
e	19,9	20,2	18,4	17,9	16,4
i	23,1	23,6	28,0	28,2	22,9
o	12,8	12,7	11,9	13,2	10,8
u	9,6	8,0	8,0	7,6	11,2
y	2,7	2,9	3,9	3,7	3,4
ä	12,8	12,5	10,7	10,5	8,8
ö	0,6	1,0	0,8	1,2	1,1

Myös toinen vokaalien taajuuksia koskeva piirre näkyy jo taulukoista 2 ja 3: itämurteissa *e*:n frekvenssi nousee kauttaaltaan *a*:n tasalle, jopa ohikin. Havainnollistaakseni asiaa julkaisen taulukon 5, josta näkyvät vokaalien keskinäiset frekvenssit ja prosenttiosuudet. Kun lukuja arvioidaan tarkemmin, näyttää siltä, että SM:ssa *e* valtaa tilaa ensisijaisesti *i*:ltä ja tämän taajuuskorotuksen avulla tavoittelee *a*:n asemia. Yhtenä selityksenä on diftonginreduktio. KM:ssa *i*:n frekvenssi taas on aivan huippuluvuissa, jopa yli 28 prosentin, mutta *e* pystyy silti kilpailemaan tasapäisesti *a*:n kanssa, jonka taajuusluvut täällä ovat alempana kuin missään muussa murteessa (kKM:ssa jopa alle 18 %:n).

[3] Kuten aiemmin totesin, murreaineistoni edustaa pääfunktioltaan kertovaa ilmaisua. Tällöin on ilmeistä, että imperfektimuotoja on aineistossa jonkin verran tavanomaista runsaammin; näin ollen imperfektin *i*:n tiheä esiintyminen voi nostaa *i*:n prosenttilukuja. Usein tämä *i* on kuitenkin myös heittyneenä (*rupes*, *anto tulla*), eikä tämän erillisen piirteen vaikutus voine kuitenkaan olla kovin merkittävä. Varsin runsaasti imperfektimuotoja sisältynee myös moneen *a*-voittoiseen otokseen (esim. sanomalehtiuutiset, kansansatu, Linnan romaani ym.).

6. Tilastojen vertailua

Suppeahkoon aineistooni perustuva grafeemitilastoni käy erittäin hyvin yksiin Pääkkösen julkaiseman mammuttiaineistoon perustuvan kirjakielen grafeemitilaston kanssa: ensimmäinen ero on se, että listassani kahdeksannella sijalla oleva *u* on Pääkkösen luettelossa vasta kymmenentenä (silti prosenttiluvut täsmäävät varsin hyvin: 5,3/5,06). Asetan vielä tähän oman tilastoni ja Pääkkösen kuusi ensimmäistä grafeemia prosenttilukuineen rinnakkain: *a* 12,1/11,90; *i* 10,9/10,64; *t* 9,6/9,77; *n* 8,9/8,67; *e* 7,83/8,21; *s* 7,79/7,85.

Viimeisenä julkaistava taulukko nro 6 kertoo vokaalien ja konsonanttien useuksien suhdeluvut eri murteissa, murrepuhunnassa yleensä ja kirjakielessä. Taulukosta nähdään, että päämurrealueista vain LM:n vokaalien frekvenssi jää 46 prosenttiin; toiseksi pienin taajuus on jo lähes 47,5. Omalla tavallaan tämäkin korostaa LM:n erikoisasemaa suomen murteiden joukossa. Toisaalta huomataan, että itämurteet ja erityisesti kaakkoismurteet ovat suomen vokaalivaltaisimpia murteita. Tuskin tämäkään on varsinainen yllätys; merkittävää on kuitenkin todeta, että kaakkoismurteissa vokaalien suhdeluku kohoaa jopa 49,8—49,9:ään — ollaan siis jo aivan lähellä vokaalien ja konsonanttien välistä »ihanteellista tasapainoa».

TAULUKKO 6. Vokaalien ja konsonanttien suhde murteissa ja kirjakielessä.

	Vokaalit	Konsonantit		Vokaalit	Konsonantit
LM	45,99	54,01	leSM	48,99	51,01
LsM	48,28	51,72	ipSM	48,42	51,58
pHM	47,73	52,27	eKM	49,79	50,21
eHM	47,38	52,62	kKM	49,86	50,14
ekPM	48,59	51,41	*Murteet*:	*48,22*	*51,78*
pP&PpM	48,33	51,67	*Kirjakieli*:	*47,64*	*52,36*

Taulukosta 6 nähdään vielä, että murteissa ja kirjakielessä ei näy vokaalien ja konsonanttien suhdeluvun välillä merkittävää eroa: suhde on suunnilleen 48 : 52. Tämä pitää hyvin tarkasti yhtä esimerkiksi Häkkisen ja Pääkkösen aikaisempien tulosten kanssa.

Silti taulukosta on pääteltävissä, että murteessa — ja yleensä puhekielessä? — on pyrkimys kirjakieltä vokaalivaltaisempaan ilmaisuun. Lieneekin niin, että myös foneemi- ja grafeemitaajuuksien valossa yleispuhekieli asettuu välittäväksi kielimuodoksi murrepuhunnan ja kirjakielen välille[4].

[4] Kirjoituksen latomisen jälkeen on ilmestynyt Anneli Pajusen ja Ulla Palomäen Tilastotietoja suomen kielen rakenteesta 1, joka sivuaa tässä esitettyä.

Liite

Murreaineistoni on murrealueittain ja pitäjittäin lueteltuna seuraava. Otoksen koko pitäjää kohti on siis noin 10 000 foneemia.

Lounaismurteet (LM): (a) luoteisryhmästä 003/1 Rauma ja 017 Kalanti, (b) kaakkoisryhmästä 064 Perniö ja 075 Paimio sekä (c) Turun lähistöä ja rannikkomurteita edustavina 046 Masku ja 055/1 Rymättylä. LM:sta päädyin kolmen ryhmän valintaan osittain osajoukkoja vertaillakseni, osittain siksi, että tältä alueelta oli tuolloin runsaasti aineistoa saatavissa. Tästä johtuu, että aineistoni lopullinen foneemimäärä on 400 000:n sijasta lähellä 420 000:ta; kun lisäksi ottaa huomioon katkaisematta jätettyjen saneiden vaikutuksen, tuli lopulliseksi foneemimääräksi 420 100.

Lounaiset siirtymämurteet (LsM): 101 Merikarvia, 126 Vampula, 131 Loimaa ja 161 Somero.

Pohjoishämäläiset murteet (pHM): (a) 211 Kankaanpää ja 231 Punkalaidun; (b) 255 Pirkkala ja 275 Juupajoki.

Etelähämäläiset murteet (eMH): (a) 305 Sääksmäki ja 322 Renko; (b) 332 Lammi ja 364 Askola.

Etelä- ja keskipohjalaiset murteet (ekPM): (e) 413 Laihia ja 425 Ylihärmä; (k) 453 Toholampi ja 498 Kestilä.

Pohjoispohjalaiset ja *Peräpohjolan murteet* (pP&PpM): (pP) 504 Paavola ja 522 Yli-Ii; (Pp) 542 Kemi ja 545 Sodankylä.

Läntiset ja eteläiset savolaismurteet (leSM): (a) Sysmä ja 646 Laukaa; (b) 686 Mikkeli ja 693 Punkaharju.

Itäiset ja pohjoiset savolaismurteet (ipSM): (a) 716 Rantasalmi ja 737 Nilsiä; (b) 763 Posio ja 772 Kitee. Savolaismurteista ei tuolloin juuri ollut tiedostoja käytettävissä; tämä selittää sen, että jouduin sijoittamaan esim. Posion ja Kiteen samaan ryhmään. Molemmat sentään edustavat suhteellista itäisyyttä.

Kaakkoismurteet, Etelä-Karjala (eKM): (a) 802 Savitaipale ja 807 Ruokolahti; (b) 811 Luumäki ja 821 Nuijamaa; Kannaksen Karjala (kKM): (a) 861 Rautu ja 865 Räisälä; (b) 874 Lumivaara ja 882 Sortavala. Esim. savolaismurteisiin verrattuna on kaakkoismurteista otettu varsin pieneltä alueelta suhteellisen suuri otos. Tätä kuitenkin puoltaa näiden murteiden heterogeenisuus ja erityisesti pyrkimykseni tasapainoiseen otokseen päämurrealueiden ja itä- ja länsimurteiden kesken.

Kirjakielen aineistoni sisältää 10 000 kirjaimen otokset seuraavista neljästä tekstilajista: 950 sanomalehtien pääkirjoitukset; 953 sanomalehdet: uutiset; 961 aikakauslehtien reportaasit; 972 historian tietokirjat.

LÄHTEET

HAKULINEN, LAURI 1938: Mikä on luonteenomaista suomen kielen äännerakenteelle? — Virittäjä 42 s. 269—280.
———— 1941, 1968, 1979: Suomen kielen rakenne ja kehitys. 1. painos, I osa; 3. painos; 4. painos. Helsinki — Keuruu.
HÄKKINEN, KAISA 1977: Tilastotietoja suomen kielen äännerakenteesta. — Sananjalka 19 s. 57—68.
KARLSSON, FRED 1969: Suomen yleiskielen segmentaalifoneemien paradigma. — Virittäjä 73 s. 351—362.
KOHONEN, VILJO—SALMELA, JUSSI 1978: Aineiston valinnan ja automaattisen tietojenkäsittelyn ongelmia kielitieteellisessä tutkimuksessa. Working Papers on Computer Processing of Syntactic Data (toim. Erik Andersson). Publications of the Åbo Akademi Research Foundation 38. Åbo.
LEPISTÖ, EINO 1938: Vampulan murteen äänteiden yleisyystilastoa. — Virittäjä 42 s. 45—54.
MIKKONEN, VALDE 1972: Suomen kielen kirjainten frekvenssi- ja informaatio-ominaisuuksia. — Acta Botnica 1972, Turun yliopiston pohjalaisen osakunnan vuosikirja, s. 20—39. Turku.
PESONEN, JAAKKO 1971: Sanamuodot ja niiden kirjainrakenne suomenkielisessä sanomalehtitekstissä. — Research reports n:o 6/1971, Department of Special Education, University of Jyväskylä.
PÄÄKKÖNEN, MATTI 1973: Tilastotietoja suomen yleiskielen grafeemeista. — SUSA 1973 s. 318—322.
RUOPPILA, VEIKKO 1936: Äänteiden yleisyystilastoa Lemin murteesta. — Virittäjä 40 s. 127—131.
SETÄLÄ, VILHO 1972: Suomen kielen dynamiikkaa. Suomi 116:3. Suomalaisen Kirjallisuuden Seura. Helsinki.

On the frequencies of Finnish phonemes and graphemes

OLLI JÄRVIKOSKI

Because Finnish orthography is phonemically very precise, in practice the concepts of the phoneme and the grapheme are very close. Whether we count sounds or letters, phonemes or graphemes, the basis of the calculation remains more or less the same, and so the results for different periods and different materials are fairly comparable. The article looks at the frequency counts for Finnish at different times, and compares them with each other and the writer's own research.

The oldest phoneme frequency counts date from the 1930's, and the only ones intended to be generalizable for the whole language are those of Lauri Hakulinen (1938). According to these results the most frequent sound was *i*, followed by *t*, *a*, *e*, *s* and *n*. Hakulinen repeated these figures as such in 1968. In the 1970's several scholars (Pesonen 1971, Mikkonen 1972, Setälä 1972, Pääkkönen 1973 and Häkkinen 1977) studied phoneme and grapheme frequencies with samples of various sizes and types, and using different terminologies. On the basis of these studies Hakulinen concluded in 1979 (SKRK = 'The structure and development of the Finnish language', 4th edition) that the most frequent phoneme (and grapheme) in Finnish is *a*.

The present writer has arrived at the same result for *graphemes*: *a* is the most common. However, on the basis of an extensive sample (400,000 phonemes) covering different Finnish dialects, the most frequent *phoneme* in Finnish turns out to be *i*, both in each dialect individually (see tables 2 and 3) and in the dialects as a whole (table 4, phonemes; compare table 4, graphemes). The most important explanatory factor for this difference is probably the fairly frequent

occurrence of apocope in Finnish dialects. In Standard Finnish *a* is a very frequent grapheme in word-final position (a frequency of 25.2 %), but in the dialects the phoneme *a* at the end of a word is noticeably less frequent (16.8 %). On the other hand word-final *i* is more common in the dialects (13.9 %) than in the standard language (10.3 %).

The order of the six most frequent graphemes is *a*, *i*, *t*, *n*, *e*, *s*, which corresponds closely to earlier counts for the standard language. The six most frequent phonemes are *i*, *t*, *a*, *e*, *s*, *n*, which by chance agrees exactly with Hakulinen (1938), although Hakulinen's old material probably did not include dialects (see table 1).

Table 2 shows the frequency order for phonemes by dialect group, and table 3 the phoneme frequencies as percentages in ten dialect areas. In table 4 the dialect phoneme frequencies are compared with the grapheme frequencies for the standard language, as percentages. Table 5 gives detailed percentages for the distribution of vowels in different dialects (LM—kKM) and in the standard language. The comparison shows that in the eastern dialects (SM and KM) the frequency of *e* is as great or greater than that of *a*. Table 6 gives the ratio of vowels to consonants as percentages first in different dialects (LM—KM), then in the whole of the dialect material, and lastly in the standard language. The south-eastern dialects of south Karelia (eKM—kKM) turn out to be the most vowel-dominant, with a vowel—consonant ratio of practically 50 %—50 %. The most consonant-dominant dialects are the south-western ones (LM). Taken as a whole, the dialect material is only slightly more vowel-dominant than the standard language material.

From the sound-structure point of view, therefore, easily the six most important sounds (in all dialects and also in the standard language) are the vowels *a*, *e* and *i* and the consonants *n*, *s* and *t*. Another group of 12 phonemes and graphemes is even more clearly distinguished when we add the vowels *o*, *u* and *ä* and the consonants *k*, *l* and *m*. It is on these six vowels and six consonants that the sound structure of Finnish basically rests. The fact that in this group of twelve the vowels and consonants are equally balanced is a reflection of the stable relationship between vowels and consonants in the language as a whole.