KAISA HÄKKINEN

Tilastotietoja suomen kielen äännerakenteesta

Tunnetuimmat suomen kielen äännerakenteen luonteenomaisia piirteitä valaisevat numerotiedot ovat Lauri Hakulisen Suomen kielen rakenne ja kehitys -teoksen äänteiden suhteellisia frekvenssejä kuvaavat luvut (Hakulinen 1968, 15—19). Luvut pohjautunevat Hakulisen omiin laskelmiin siitä päätellen, etteivät ne käy yksiin mainitun artikkelin alaviitteissä lueteltujen suomen kielen muitten äännetilastojen kanssa. Laskentaperusteista, materiaalin koostumuksesta ja laajuudesta Hakulinen ei anna mitään tietoja. Samat luvut on esitetty jo aikaisemmin vuoden 1938 Virittäjässä (Hakulinen 1938, 269—280). Kun laskelmat ovat näin vanhoja, on ilmeistä, että ne on suoritettu käsin eikä materiaali ole voinut olla kovin laaja. Nykyään vastaavien laskelmien suorittaminen suuremmista korpuksista on ATK-menetelmien ansiosta tullut käytännössä mahdolliseksi, ja näin ollen pidän tarpeellisena suorittaa vertailua Hakulisen esittämien lukujen ja laajemman, n. 140 tekstisivua sisältävän materiaalin perusteella laskettujen arvojen välillä, kun mainitun aineiston fonotaktisen tutkimuksen ohessa tarjoutui tilaisuus tällaisia laskelmia melko vaivattomasti suorittaa. Tämän tutkimuksen ATK-teknisestä puolesta vastaa Turun yliopiston fonetiikan oppiaineen teknikko M. Mattila. Vertailun helpottamiseksi noudattelen artikkelissani Hakulisen käyttämää esitysjärjestystä.

Kvantitatiivisen kielentutkimuksen asiantuntijat ovat havainneet, että äännesuhteet säilyvät hämmästyttävän vakioina tekstin laadusta riippumatta (vrt. esim. Karlgren 1968, 142; vrt. myös Setälä 1972, 17). Näin ollen pidän mahdollisena vertailla Hakulisen esittämiä lukuja omiin laskelmiini. Aineistoni muodostuu suomalaisesta kansansatusovitelmasta (Lehtonen 1953). Päädyin tähän korpukseen, koska pyrin mahdollisuuksieni mukaan välttämään nuoria lainasanoja sisältäviä tekstejä, joissa esiintyy runsaasti supisuomalaiselle äännesysteemille vieraita konsonantteja ja äännekombinaatioita. Materiaalia reikäkorteille lävistäessäni jätin pois kaikki erisnimet ja interjektiot, ja työn päät-

tyessä kävi selville, ettei aineistoon jäänyt yhtään sellaista äännettä, jota ei voisi esiintyä vanhassa omaperäisessä sanassa. Fonotaksin kohdalla nuorehkoja piirteitä tuli välttämättä mukaan jonkin verran (esim. sananalkuisia konsonanttiyhtymiä: *kr*uunu, *pr*inssi), mutta tällaisten tyyppitapausten marginaalisuus kuvastuu suoraan niitten frekvenssissä. Yksi korpustutkimuksen periaatteistahan on, ettei mitään suljeta pois ennakolta, vaan kaikki tutkittavaan alaan kuuluva materiaali otetaan mukaan ja käsitellään samanveroisena. Tässä kohdin yllä mainitut erisnimet ja interjektiot voidaan käsittää suomen kielen äännerakenteessa omiksi erikoissysteemeikseen, koska niitten kohdalla rakenneperiaatteet ovat osittain erilaiset kuin muussa sanastossa. Erisnimissä on mm. käytössä kokonainen tavutyyppi CVVCC (esim. *Suort*tanen, Jo*taark*ka), jota yleiskielen sanastossa ei esiinny lainkaan. Interjektioissa sananalkuiset ja -loppuiset konsonanttiyhtymät ovat täysin luvallisia, jopa suosittuja (*pl*äts, *pl*umps). Koska tutkittava systeemi on suppea, tuntuu aineisto ehkä tarpeettoman suurelta (21 foneemia, n. 140 sivua tekstiä). On olemassa tapauksia, joissa otoksen suurentaminen vaikuttaa epäedullisesti lopputuloksen luotettavuuteen. Näin käy silloin, kun otos ei ole edustava. Mutta mikäli äännesuhteitten vakioisuus pitää todella paikkansa, on koko teksti välttämättä luotettavampi kuin sen osa. Vähemmälläkin materiaalilla kuin tässä käytetyllä päästäisiin varmasti yhtä luotettavaan tulokseen, mutta vaadittavan laajuuden tarkka määrittely olisi suoritettava kokeilemalla. Mainittakoon, että tekstissä esiintyvien saneiden loppuäänteitten frekvenssejä Annikki Salmelinin esittämiin vastaaviin lukuihin vertaillessani havaitsin erittäin hyvän korrelaation (Salmelin 1959, 400—403). Tästä voi päätellä, että Salmelinin käyttämä kymmenen tekstisivun laajuinen otos antaa saneiden loppuäänteitten frekvenssistä aivan yhtä luotettavaa tietoa kuin oma 140 sivun otokseni. Lävistämäni tekstiaineisto on kuitenkin koottu laajempaa fonotaksin tutkimusta varten, ja kun sen käsittely muissa vaiheissa on suoritettava yhtenä kokonaisuutena, pidin tarpeettomana vaivana käydä osittamaan sitä äännefrekvenssien tutkimusta varten. Vastaisia samantyyppisiä tutkimuksia silmällä pitäen olisi tietysti ollut hyödyllistä osittaa materiaali ja tutkia, kuinka suureen tarkkuuteen lopputuloksen osoittamien raja-arvojen suhteen erilaajuisten otoksien avulla olisi mahdollista päästä.

Vertailun kohteena olevien kahden tutkimuksen osittaisen terminologisen yhteismitattomuuden vuoksi haluan heti alussa korostaa, että olen pyrkinyt nimenomaan foneemien suhteellisen frekvenssin selvittämiseen ja uskon, että tähän Hakulinenkin on tähdännyt, vaikka hän käyttääkin koko ajan *äänne*-termiä. Varsinkin vanhemmassa kielitieteellisessä kirjallisuudessahan tämä termi on ollut kaksiselitteinen, ja

monissa yhteyksissä se esiintyy fonologisuuden suhteen indifferenttinä. Niissä tapauksissa, joissa puututaan foneettisiin seikkoihin, esim. konsonanttien tai vokaalien jaotteluun artikulaatiopaikan mukaan, on tietysti kysymys foneemien normaaleista realisaatioista, ei itse foneemeista.

Hakulisen mukaan suomen kieli on hyvin erikoislaatuinen maailman muitten kielten joukossa siinä suhteessa, että 100 vokaalia kohti käytetään 96 konsonanttia. Muissa tämän ominaisuuden suhteen tutkituissa kielissä konsonantteja käytetään enemmän kuin vokaaleja. Hakulinen huomauttaa kuitenkin itse, etteivät vertailtavat luvut välttämättä ole täysin luotettavia erilaisten laskentaperusteitten ja otoslaajuuksien vuoksi. Lähteet ovat vanhoja, ja kun muistamme, että johdonmukainen fonologinen ajattelutapa alkoi vasta kehittyä 1900-luvun alussa, on pidettävä ilmeisenä, että joissakin tapauksissa on yksinkertaisesti laskettu grafeemeja. Mitään virhemarginaaleja ei ole ilmoitettu, ja kun esim. suomen kohdalla vokaalien ja konsonanttien suhde on niinkin tasainen kuin 100/96, on kyseenalaista, voidaanko vokaalien suurempaa taajuutta korostaa siten kuin Hakulinen on tehnyt. Omat laskelmani osoittavat suhteen päinvastaiseksi, niitten mukaan koko tekstimateriaalin foneemistosta 52 % on konsonantteja, 48 % vokaaleja. Jäännöslopuke on sekä Hakulisen laskelmissa että omissani jätetty kokonaan pois (jäännöslopukkeen fonologisesta asemasta ks. esim. Karlsson—Lehtonen 1977, 61—64). Puhutussa kielessä konsonanttien osuus on siis vielä suurempi, ja se kasvaa edelleen, jos yleiskielestä siirrytään arkikieleen, jossa sananloppuisten konsonanttien osuus loppuheiton vuoksi on huomattavasti suurempi kuin kirjoitetussa kielessä. Yllä mainittujen lukujen perusteella voidaan nähdäkseni vain todeta, että konsonanttien ja vokaalien suhde suomen kielessä näyttää hyvin tasaiselta, mikä tietysti on huomionarvoista sinänsä. (Vrt. myös Setälä 1972, 40.)

Suomen kahdeksan vokaalia voidaan esittää yleisyysjärjestyksessä seuraavasti:

Hakulinen	Häkkinen
1. i 27 %	1. a 26 %
2. a 23 %	2. i 24 %
3. e 16 %	3. e 15 %
4. o 10 %	4. ä 12 %
5. u 10 %	5. u 10 %
6. ä 9 %	6. o 9 %
7. y 3 %	7. y 3 %
8. ö 1 %	8. ö 1 %
Σ 99 %	Σ 100 %

Σ 99 % Σ 100 %:n asemesta johtuu desimaalilukujen pyöristämisen

aiheuttamasta epätarkkuudesta.

Listat ovat täysin yhtäpitävät kolmessa kohdin: *u*:n, *y*:n ja *ö*:n sijoituksen ja prosentuaalisen osuuden suhteen. Eroavuuksista herättää huomiota *i*:n ja *a*:n keskinäinen järjestys sekä *ä*:n sijoittuminen *o*:n ja *u*:n suhteen. Absoluuttisten lukujen valossa *a*:n ja *i*:n suhde otoksessani on 20648:19102, ts. *a*:ta oli 1546 kappaletta enemmän kuin *i*:tä. Tämä 1546 on enemmän kuin koko *ö*-vokaalin absoluuttinen frekvenssi (714) ja n. 2 % koko materiaalin sisältämien vokaalien määrästä. Mainittakoon, että myös Salmelin päätyy saneiden loppuäänteiden frekvenssejä tarkastellessaan tulokseen, että *a* on suomen yleisin vokaali (Salmelin 1959, 402), samoin Vihtori Peltonen kirjapainojen käyttämiin kirjasinmääriin pohjautuvissa laskelmissaan (Peltonen 1926, 91) ja Vilho Setälä Uuden Testamentin suomennoksen (1913) äännefrekvenssitutkimuksessaan (Setälä 1972, 17). χ^2-testin avulla on mahdollista arvioida, voivatko yllä esitetyt listat yleensä olla saman jakauman kuvaajia, joitten erot perustuvat vain sattumaan. Merkitsemällä omat lukuni teoreettisiksi arvoiksi ja Hakulisen luvut havaituiksi frekvensseiksi olen laskenut χ^2:n arvoksi kaavan

$$\chi^2 = \Sigma \frac{o_j^2}{e_j} - N \qquad \text{(Spiegel 1972, 201)}$$

mukaan 1,95. Tämä luku alittaa selvästi χ^2:n kriittiset arvot merkitsevyystasoilla 99 %, 95 % ja 90 % (= 18,5, 14,1 ja 12,0; $\nu = 7$, Spiegel 1972, 345). Sarjojen välillä vallitsee siis vahva korrelaatio ja ne voivat ilmeisesti kuvata samaa jakaumaa. Mutta kuinka luotettavina voimme pitää mainittuja prosenttilukuja? Oma otokseni sisältää 80476 vokaalia. Tämä merkitsee sitä, että prosenttiluvun muutos yhdellä yksiköllä edellyttää n. 805 kappaleen muutosta absoluuttisessa frekvenssissä. Jos otos sen sijaan sisältäisi vain 1000 foneemia, olisi yksi prosentti tästä 10. Koska Hakulisen otoksen täytyy olla pienempi kuin omani, siinä sattuman osuus on todennäköisesti suurempi. Jos ajattelemme oman otokseni esim. 80:nä tuhannen foneemin otoksena ja muistamme, että sattuman aiheuttama poikkeama voi yhtä suurella todennäköisyydellä vaikuttaa kahteen vastakkaiseen suuntaan, käy selväksi, että nämä vastakkaiset poikkeamat todennäköisesti eliminoivat toistensa vaikutuksen. Jos siis merkitsemme tietyn suuntaisen poikkeaman todennäköisyyden 1/2:ksi, on todennäköisyys, että kahdesta poikkeamasta molemmat ovat samansuuntaisia vain 1/4 ja todennäköisyys, että kolme poikkeamaa ovat samansuuntaisia 1/8 jne. Yllä olevasta käynee ilmi, että kahden kovin erilaajuisen otoksen perusteella laskettuja prosenttilukuja voi verrata toisiinsa vain varauksin. Jos haluaisimme absoluuttisten

lukujen valossa tarkastella, millainen tulos omasta laskelmastani olisi pitänyt saada, jotta sen prosenttiluvut täsmäisivät Hakulisen esittämien lukujen kanssa, havaitsisimme, että *a*-vokaalia olisi 20648 kappaleen sijasta pitänyt olla 18509 ja vastaavasti *i*:tä 19102:n asemesta 21729. Mikäli Hakulisen otos olisi ollut yhtä laaja kuin omani ja jos hän olisi todella saanut sen perusteella yllä kuvatun kaltaiset absoluuttiset frekvenssit, täytyisi eroa pitää jo hyvin merkitsevänä edellä kuvatun testin perusteella.

Taka- ja etuvokaalien suhde on Hakulisen mukaan 43:57. Tämä ei anna minkäänlaista kuvaa siitä, kuinka sanasto jakautuu taka- ja etuvokaalisuuden suhteen. Tiedämmehän, että suomen etuvokaaleista kaksi, *e* ja *i*, ovat erikoisasemassa sikäli, että ne voivat kombinoitua sekä taka- että etuvokaalien kanssa. Tämän vuoksi *e*:tä ja *i*:tä voi nimittää etuvokaaleiksi vain foneettisessa mielessä, fonotaktisesti ne ovat neutraaleja. Koska sanojen rakentumisperiaatteissa ovat määräävinä monet muutkin seikat kuin vokaalien absoluuttiset esiintymistodennäköisyydet, on mahdotonta määritellä todennäköisiä kombinoitumisfrekvenssejä pelkkien lukumääräsuhteitten perusteella. Jonkinlaista arviointia voi kuitenkin suorittaa olettamalla, että neutraaleista vokaaleista "takavokaalisten", ts. takavokaalisissa sanoissa esiintyvien osuus on suurin piirtein sama kuin puhtaitten takavokaalien osuus vokaalien kokonaismäärästä. Takavokaalien (*a*+*o*+*u*), Neutraali- (*e*+*i*) ja etuvokaalien (*y*+*ä*+*ö*) suhde laskelmissani on 45:39:16. Nyt siis 39 on jaettava edelleen tässä samassa suhteessa. Näin saadaan taka- ja "takais-neutraali"-vokaalien yhteenlasketuksi osuudeksi 45 + 18 (18 = n. 45 % 39:stä) = 63 %, etisten ja "etis-neutraaleitten" summaksi tulee 16 + 6 = 22 % ja puhtaasti neutraaleitten vokaalien osuudeksi jää 39—18—6 = 15 %. Laskin satunnaisesti valitusta 263 sanan otoksesta (yksi tekstisivu) eri tyyppisiä etu/neutraali/takavokaalikombinaatioita sisältävät sanat ja sain seuraavan tuloksen:

	taka	taka + neutr.	neutr.	etu + neutr.	etu
otos	27 %	33 %	11 %	23 %	6 %
		60 %		29 %	
teoreett.		63 %	15 %	22 %	

Prosenttiluvuille ei sinänsä voi antaa kovin suurta arvoa, mutta molemmista sarjoista käy selvästi ilmi, että tekstissä tuntuu takavokaaleja sisältävien sanojen osuus suurempana kuin etuvokaaleja sisältävien sanojen osuus. Tämä suhde on juuri päinvastainen kuin se, mihin voisi päätyä ottamalla huomioon foneettisesti etisten ja takaisten vokaalien lukumääräsuhteet. Tämän avulla voidaan ehkä selittää myös suomen kielen takaista artikulaatiobaasista: On todennäköisempää, että sanassa

on takavokaaleja, kuin se, ettei niitä siinä ole. Tämän vuoksi on yleensä taloudellisempaa pitää kielen asento takaisena artikulaatiota aloitettaessa.

Samaan tavuun kuuluvien VV-sekvenssien sekä lyhyitten vokaalien lukumääräsuhde on Hakulisen mukaan 1:3,2. Oma laskelmani antoi tulokseksi 1:3,3, mitä on käytännössä pidettävä samana.

Hakulinen esittää myös suomen yksitoista yleisintä äännettä. Tähän listaan sisältyy muuan ongelma. Kun mukana on *n*, joka suomen kirjainjärjestelmässä on kaksiselitteinen, olisi välttämätöntä selvittää, onko frekvenssi laskettu mekaanisesti *n*-grafeemien perusteella vai onko luvusta poistettu ne tapaukset, joissa *n* esiintyy *k*:n tai *g*:n edellä ts. on ainakin foneettisesti /ŋ/. Tavunloppuisen /ŋ/-äänteen fonologinen tulkinta ja siten myös /*n*/:n ja /ŋ/:n frekvenssi riippuu siitä, millainen foneemikäsitys valitaan. On joka tapauksessa selvä, että suomen kielessä on /ŋ/-foneemi, tämä käy ilmi esim. minimitripletistä *ramman* — *rannan* — *raŋŋan*. Mutta mikä on /ŋ/:n distribuutio? Puhtaasti fysikaalisen foneemikäsityksen mukaan voimme käyttää vain foneettisia kriteerejä. Jos kielessä on /ŋ/-foneemi, on jokainen samat foneettiset piirteet sisältävä äänne katsottava /ŋ/-foneemiin kuuluvaksi. Esim. tapauksissa *pojaŋkin* tai *pojaŋko* täytyy *k*:n edellä oleva nasaali katsoa fonologisesti /ŋ/:ksi, koska se sisältää samat foneettiset tuntomerkit kuin /ŋ/*raŋŋan*-tapauksessa. Tällaista foneemikäsitystä kieltäydytään usein hyväksymästä, koska intuitiomme *pojaŋko*-tapauksessa sanoo, että *k*:n edellä itse asiassa onkin /*n*/, joka äänneympäristönsä vuoksi on muuttunut ŋ:ksi. Toisin sanoen tuntuu paremmalta valita vaihtoehto, joka sallii tällaisen muutoksen fonologista tulkintaa konkreettisemmalla tasolla. Tämän tulkinnan intuitiivinen hyväksyttävyys näkyy siinäkin, ettemme tunne tarvetta merkitä *n*:ää millään erikoisella merkillä sen joutuessa *k*:n edelle. Kaikkiin tavunloppuisiin ŋ-tapauksiin voitaisiin soveltaa samaa periaatetta. Tämän abstraktisen eli generatiivisen tulkinnan mukaan lähdettäisiin säännöstä, että /ŋ/ voi esiintyä ainoastaan tavun alussa ja sillä on foneettisella tasolla sama assimiloiva vaikutus edeltävään nasaaliin kuin *k*:llakin. Suomen kirjoitusjärjestelmä noudattaa juuri tätä periaatetta; tosin ŋ-merkki on korvattu *g*:llä, joka nuorten lainasanojen myötä on tullut kaksiselitteiseksi. Generatiivisen tulkinnan puolesta on vielä todettava, ettei ihmisen havainnointijärjestelmä yleensäkään toimi pelkkien fysikaalisten kriteerien perusteella, vaan valintaa ja abstrahointia suoritetaan koko ajan, eikä sitä fysikaalista tarkkuutta, mihin periaatteessa olisi mahdollista päästä, läheskään sellaisenaan käytetä hyväksi. Toleranssirajojen tutkiminen olisikin erittäin mielenkiintoista eikä variaation luvallisuutta missään yhteydessä pitäisi unohtaa. Esim. fonologisen opposition puuttuminen

antaa mahdollisuudet hyvinkin vaihtelevaan realisaatioon: kun suomessa on vain yksi sibilantti, se voidaan ääntää kovin monella eri tavalla ilman että lopputuloksen ymmärrettävyys kärsii. Omissa laskelmissani olen lähtenyt siitä, että vain tavunalkuinen *ŋ* on fonologisesti */ŋ/*. Tämä tietysti edellyttää polysysteemistä foneemikäsitystä, jonka mukaan foneemijärjestelmä voi eri asemissa olla erilainen. Varmuuden vuoksi, kun en ole selvillä siitä, kuinka Hakulinen on tulkinnut nasaalin grafemaattisissa *nk-* ja *ng*-tapauksissa, esitän sekä */n/*:n että */ŋ/*:n kohdalla kaksi lukua. Näistä */n/*:n kohdalla suurempi sisältää kaikki *n*-grafeemilla merkityt äänteet; pienemmästä on vähennetty ne *n*-tapaukset, joissa nasaali on *k*:n tai *g*:n edellä, ts. on foneettisesti *ŋ*. */ŋ/*:n kohdalla vastaavasti suurempi luku sisältää kaikki foneettisin perustein */ŋ/*-fonemiin kuuluvat äänteet (= fysikaalinen tulkinta), pienempi sisältää vain tavunalkuisen */ŋ/*:n frekvenssin (= generatiivinen tulkinta). */n/*:n sijoitukseen ei tulkintaero vaikuta, sen sijaan */ŋ/* siirtyy viimeiseltä sijaltaan ylemmäs, mikäli valitaan fysikaalinen tulkinta. Esitän kaikki luvut yhden desimaalin tarkkuudella Hakulisen mallin mukaisesti. Alkuperäiset laskelmat on suoritettu neljän desimaalin tarkkuudella ja listaus kolmea desimaalia käyttäen.

Hakulinen			Häkkinen		
1.	i	12,0 %	1.	a	12,3 %
2.	t	11,5 %	2.	i	11,4 %
3.	a	10,4 %	3.	n	9,8 % (9,5 %)
4.	e	9,4 %	4.	t	8,8 %
5.	s	8,5 %	5.	e	7,4 %
6.	n	8,4 %	6.	s	7,2 %
7.	ä	5,7 %	7.	k	6,2 %
8.	l	5,7 %	8.	l	5,8 %
9.	k	5,0 %	9.	ä	5,6 %
10.	o	4,9 %	10.	u	4,9 %
11.	u	4,8 %	11.	o	4,5 %
			12.	m	2,8 %
			13.	p	2,3 %
			14.	v	2,2 %
			15.	r	2,2 %
			16.	j	2,1 %
			17.	h	2,1 %
			18.	y	1,6 %
			19.	ö	0,4 % (19. ŋ 1,0 %)
			20.	d	0,4 % (20. ö 0,4 %)
			21.	ŋ	0,1 % (21. d 0,4 %)

Hakulisen listassa ei *ä*:n sijoitus *o*:n ja *u*:n suhteen pidä yhtä pelkkien vokaalien yleisyysjärjestystä koskevan listan kanssa. Siellä *ä* oli merkitty *o*:ta ja *u*:ta harvinaisemmaksi.

Koska Hakulinen on vertaillut suomen äännefrekvenssejä mm. unkarin äänteiden yleisyysjärjestykseen ja koska unkarin osalta on saatavissa uutta tietoa, suoritan myös tässä vastaavanlaisen vertailun. Kun unkarissa pitkän ja lyhyen vokaalin vastakohta on ehdottomasti fonologinen, on väärin yhdistää lyhyitten ja kvalitatiivisesti samantyyppisten pitkien vokaalirealisaatioiden frekvenssit, kuten Hakulisen referoimassa lähteessä on tehty. Suomen ns. pitkien vokaalien toisenlainen fonologinen tulkinta ei saisi häiritä tässä kohdin, koska jokaisen kielen foneemisysteemi on selvitettävä omista lähtökohdistaan käsin. Yleensäkin olisi oltava tarkoin selvillä siitä, onko tarkoitus laskea fooneja, foneemeja vai grafeemeja. Lähteenäni olleessa artikkelissa (Kálmán 1972) on laskettu foneemien frekvenssejä, joten pidän näitä lukuja ja omiani vertailukelpoisina. Yhtenäisyyden vuoksi esitän Kálmánin luvut yhden desimaalin tarkkuudella. Lisävertailukohteeksi olen ottanut mukaan myös Sosvan vogulimurteen (= vogulin kirjakielen) foneemifrekvenssit, jotka olen laskenut Wogulische Volksdichtung -sarjan proosateksteistä poimitun otoksen perusteella laudaturtyöni yhteydessä (Häkkinen 1973). Unkarin kohdalla käytän Kálmánin tavoin oikeinkirjoitusjärjestelmän merkkejä foneemimerkkeinä, jotta muittenkin kuin unkarintaitoisten olisi helppo assosioida lista unkarinkieliseen tekstiin.

unkari (Kálmán 1972, 77)		suomi		voguli (Häkkinen 1973, 86)	
1.	e 10,9 %	1.	a 12,3 %	1.	t 10,1 %
2.	a 9,9 %	2.	i 11,4 %	2.	a 9,2 %
3.	t 7,7 %	3.	n 9,8 %	3.	i 6,9 %
4.	l 5,8 %	4.	t 8,8 %	4.	l 6,3 %
5.	n 5,5 %	5.	e 7,4 %	5.	s 6,1 %
6.	k 5,3 %	6.	s 7,2 %	6.	n 5,8 %
7.	i 4,8 %	7.	k 6,2 %	7.	m 5,2 %
8.	r 4,2 %	8.	l 5,8 %	8.	ē 5,1 %
9.	o 4,2 %	9.	ä 5,6 %	9.	ā 4,7 %
10.	m 4,1 %	10.	u 4,9 %	10.	ə 4,4 %
11.	s 3,8 %	11.	o 4,5 %	11.	β 3,8 %
12.	á 3,6 %	12.	m 2,8 %	12.	j̱ 3,0 %
13.	é 3,5 %	13.	p 2,3 %	13.	o 3,0 %
14.	g 2,5 %	14.	v 2,2 %	14.	χ 2,9 %
15.	z 2,3 %	15.	r 2,2 %	15.	γ 2,9 %
16.	d 2,1 %	16.	j 2,1 %	16.	p 2,8 %
17.	b 2,1 %	17.	h 2,1 %	17.	u 2,8 %
18.	v 2,1 %	18.	y 1,6 %	18.	r 2,8 %
19.	sz 1,9 %	19.	ö 0,4 %	19.	ō 2,6 %
20.	h 1,8 %	20.	d 0,4 %	20.	k 2,1 %
21.	j 1,8 %	21.	η 0,1 %	21.	η 1,6 %
22.	gy 1,6 %			22.	ū 1,4 %
23.	ö 1,1 %			23.	š 1,3 %
24.	ó 1,1 %			24.	kβ 1,2 %

25. f 0,9 %
26. u 0,9 %
27. p 0,8 %
28. ő 0,8 %
29. ny 0,7 %
30. cs 0,6 %
31. í 0,5 %
32. ú 0,5 %
33. ü 0,4 %
34. c 0,2 %
35. ty 0,1 %
36. ű 0,1 %
37. zs 0,1 %
38.—39. dz, dzs — %

25. ń 1,0 %
26. Î 0,8 %
27. ŧ 0,4 %
28. χβ 0,1 %

Listojen vertailu kokonaisuutena on hankalaa, koska kolmen kyseisen kielen foneemisysteemit poikkeavat toisistaan melkoisesti. Se on kuitenkin mielenkiintoista paitsi universaalien foneemistotendenssien kannalta myös siksi, että kolme mainittua kieltä ovat sukukieliä ja periaatteessa niiden kaikkien foneemisto olisi kronologisesti johdettavissa yhteisestä kantasysteemistä. Pelkkien frekvenssien perusteella emme tietenkään voi suoraan tehdä historiallisia päätelmiä, mutta joitakin vihjeitä mahdollisista kehityskuluista ne voivat antaa. Kunkin listan kahdeksalla ensimmäisellä foneemilla on vastine jokaisessa kielessä (huom! vogulin *ē*:llä ei ole lainkaan lyhyttä paria). Näistä kahdeksasta foneemista kuudella on vastine jokaisen kielen "kahdeksan kärjessä" -listalla. Tälle ei sinänsä voi panna kovin paljon painoa, koska mm. universaali merkitsemisteoria perustuu sille, että tietyt äänteet ovat yleensä kaikissa kielissä tavallisempia kuin jotkut toiset. Sen sijaan kannattaa kiinnittää huomiota tapauksiin, joissa esim. kahdessa kielessä hyvin suosittu foneemi kolmannessa onkin hyvin harvinainen. Unkarin */k/* on listassa kuudes, suomen */k/* seitsemäs, sen sijaan vogulissa */k/* on vasta 20. sijalla. Tämän perusteella on mahdollista olettaa, että vogulin */k/* on joutunut sellaisten historiallisten muutosten kohteeksi, jotka ovat vähentäneet ratkaisevasti sen frekvenssiä, esim. alkuperäinen */k/* on voinut jakautua useammaksi foneemiksi. Toisaalta unkarin */r/* on listassa kahdeksas, sen sijaan vogulin ja suomen */r/* ovat selvästi listojensa puolenvälin heikommalla puolella. Kun äännehistoriasta tiedämme, että *r*-äänne on yleensä kaikissa kielissä säilynyt melko hyvin, emme voi oikeastaan olettaa, että suomen ja vogulin */r/* olisivat jostakin syystä menettäneet asemiaan vaan päin vastoin, että unkarin */r/* on tullut suosituimmaksi. Tähän voisivat olla syynä lainasanat. Koska turkkilaisperäisessä sanastossa *r*-äänteen osuus silmämääräisesti arvioiden tuntui suurelta, laskin M. Räsäsen turkkilaiskielten etymologisessa sanakirjassa esittämät unkarin (mah-

dollisesti) turkkilaisperäiset sanat. Näitä oli kaikkiaan 240, ja peräti 98 niistä sisälsi ainakin yhden /r/-foneemin. (Vrt. Räsänen 1971, 119—120.)

Dentaali-, labiaali- ja palataalikonsonanttien frekvenssejä tutkiessaan Hakulinen toteaa näitten suhteeksi mainitussa järjestyksessä 4:1:1. Lisäksi hän mainitsee, että useissa muissa Euroopan kielissä, mm. unkarissa, vastaava suhde on 6:2:1. Eron hän toteaa johtuvan siitä, että suomessa dentaalit ja labiaalit ovat suhteellisesti harvinaisempia kuin muissa kielissä. Tätä olettamusta on syytä tarkastella hieman lähemmin.

Omien laskelmieni mukaan dentaali-, labiaali- ja palataalikonsonanttien suhde suomen kielessä on 5:1,5:1. Konsonanttiemme lukumäärät artikulaatiopaikan mukaan jakaen ovat 6:3:4 (= 13 kons.). Unkarissa dentaaleja on 16, labiaaleja 5 ja palataaleja 4 (= 25 kons.) (ks. esim. Keresztes 1974, 19—22[1]). Kálmánin taulukon perusteella on mahdollista esittää vastaava dentaali-, labiaali- ja palataalikonsonanttien suhde laskemalla yhteen kuhunkin artikulaatiopaikkaluokkaan kuuluvien äänteiden frekvenssi, ja tulokseksi saadaan, että sadasta äänteestä konsonantteja on todennäköisesti 58, näistä edelleen dentaaleja 37, labiaaleja 10 ja palataaleja 11. Tämä merkitsee siis suhdetta 37:10:11 eli karkeasti 4:1:1. Näin saatujen suhdelukujen perusteella suomen konsonanttisuhteet eivät vaikuta kovinkaan erikoislaatuisilta. Artikulaatiopaikkaluokkiin kuuluvien äänteiden absoluuttinen lukumäärä ei näköjään heijastu suoraan luokan suhteellisessa frekvenssissä. Tämä tulee havainnollisesti näkyviin, jos tutkimme erityisesti dentaalien osuutta konsonanttien kokonaismäärästä suomessa ja unkarissa. Suomen osalta esittämäni suhdeluku 5:1,5:1 merkitsee käytännössä sitä, että 5/7,5 konsonanteista on dentaaleja (5 + 1,5 + 1 = 7,5), unkarissa taas dentaalien osuus on 4/6 (4 + 1 + 1 = 6). Jos sievennämme suhteet, saamme molemmissa tapauksissa tulokseksi, että dentaaleja on 2/3 eli täsmälleen yhtä paljon. Kannattaa huomata, että myös Hakulisen unkarin osalta esittämä suhdeluku 6:2:1 merkitsee samaa (6/9 = 2/3). Pelkkien suhdelukujen vertaileminen keskenään on siis hyvin epähavainnollista, koska eri sarjojen yksittäiset luvut eivät ole keskenään vertailukelpoisia. Näin ollen olisi varminta tyytyä prosenttilukuihin, jolloin suhteet voidaan myös ilmaista hieman tarkempina. Näin laskien saadaan seuraavat sarjat:

[1] En viittaa tässä kohdin Kálmánin artikkeliin siitä syystä, että siinä ei ole erotettu palataalisia ja palataalistuneita äänteitä toisistaan.

	dentaaleja	labiaaleja	palataaleja
käyttöfrekv.			
suomi	66 %	14 %	20 %
unkari	64 %	17 %	19 %
absol. lukum.			
suomi	46 %	23 %	31 %
unkari	64 %	20 %	16 %

Näitten sarjojen perusteella näyttää siltä, että suomenkielisessä tekstissä dentaalit ovat suhteellisesti jopa hieman suositumpia kuin unkarissa. Jotta päästäisiin tutkimaan universaaleja tendenssejä, olisi vertailukohtia saatava lisää. Tämänkin taulukon perusteella voidaan kuitenkin jo olettaa, että konsonanttien suhteellinen jakautuminen tekstissä eri artikulaatiopaikkaluokkien kesken on suurin piirtein vakio kielestä ja äänteiden absoluuttisesta lukumäärästä riippumatta.

Avo- ja umpitavujen määrän suhteeksi Hakulinen ilmoittaa 1,11:1 ts. avotavuja käytetään hieman enemmän kuin umpitavuja. Olen päätynyt periaatteessa samaan tulokseen, joskin lukuni korostavat hieman enemmän avotavujen suosituimmuusasemaa, 1,38:1. Tässä, kuten muissakin esitetyissä tilastoissa, on selvästi havaittavissa, että Hakulisen esittämät tulokset käyvät yhteen omieni kanssa sitä paremmin mitä harvempia muuttujia kerrallaan tarkastellaan. Toisin sanoen Hakulisen käyttämä materiaali on joissakin tapauksissa riittävä, toisissa liian suppea.

Lähteet:

Hakulinen 1938: Lauri Hakulinen: Mikä on luonteenomaista suomen kielen äännerakenteelle? Virittäjä 1938, s. 269—280. Helsinki. | *Hakulinen 1968:* Lauri Hakulinen: Suomen kielen rakenne ja kehitys. Kolmas, korjattu ja lisätty painos. Keuruu. | *Häkkinen 1973:* Kaisa Häkkinen: Vogulin Sosvan murteen äännerakenteesta. Suomalais-ugrilaisen kielentutkimuksen laudaturtyö (painamaton). Helsinki. | *Kálmán 1972:* Béla Kálmán: Hungarian historical phonology. Loránd Benkő — Samu Imre: The Hungarian Language, s. 49—84. Budapest. | *Karlgren 1968:* Hans Karlgren: Statistical methods in phonetics. Bertil Malmberg: Manual of Phonetics s. 129—154. Amsterdam. | *Karlsson—Lehtonen 1977:* Fred Karlsson — Jaakko Lehtonen: Alkukahdennus. Näkökohtia eräistä suomen kielen sandhi-ilmiöistä. Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja 2. Turku. | *Keresztes 1974:* László Keresztes: Unkarin kieli. Pieksämäki. | *Lehtonen 1953:* Joel Lehtonen: Tarulinna. Kansansatusovitelmia Suomen lapsille. Helsinki. | *Räsänen 1971:* Martti Räsänen: Versuch eines etymologischen Wörterbuchs der Türksprachen II. Wortregister; zusammengestellt von István Kecskeméti. Lexica Societatis Fenno-Ugricae XVII, 2. Helsinki. | *Salmelin 1959:* Annikki Salmelin: Kirjasuomen saneen loppuäänteiden yleisyystilastoa. Virittäjä 1959, s. 400—403. Helsinki. | *Setälä 1972:* Vilho Setälä: Suomen kielen dynamiikkaa. Suomi 116, 3. Helsinki. | *Spiegel 1972:* Murray R. Spiegel: Theory and Problems of Statistics. New York.

KAISA HÄKKINEN: *Statistical information about the phonemic structure of Finnish*

In order to get reliable statistical information it is necessary that the material should be sufficiently extensive and representative. It has been shown by specialists in the quantitative study of language that the nature of a text has no particular effect on the relative frequency of phonemes. I have therefore made a statistical study based on an adaptation of folk tales, approximately 140 pages in length (Lehtonen, 1953), using the ADP methods.This I compare with the figures given at the beginning of Lauri Hakulinen's work *The Structure and Development of the Finnish Language* (Hakulinen, 1968, pp. 15—19). The results are dissimilar, the most striking difference being the order of the most common Finnish phonemes. In Hakulinen's list *i* is in the first place (27 % of the vowels and 12 % of all the phonemes). According to my own calculation by far the commonest is *a* (26 % of the vowels and 12.3 % of the totality). I find *i* to be in second place (24 % of the vowels and 11.4 % of the totality), whereas in Hakulinen *a* is the second commonest vowel (23 %) and is third in the totality (10.4 %) He puts *t* before this (11.5 %) and in my list it comes only in fourth place (8.8 %). In a study based on random selection it is rare, even suspicious, if there is absolute agreement between the results of different calculations. The percentage differences may not seem very great, but when we take into account the extent of my material and the absolute frequencies of the phonemes, the discrepancy is significant and the explanation may lie in the smallness and one-sidedness of Hakulinen's material. It is true that the scope of his material is not mentioned, but since the calculations were achieved manually it cannot be very great. On the other hand 140 pages is certainly unnecessarily large, but this cannot affect the final result negatively if the theory about the regularity of occurrence of phonemes is valid.

I have also compared the phoneme frequency of Finnish, Hungarian and Vogul. This is a difficult overall undertaking because the phonemic systems are different. It is, however, interesting to note that the eight commonest phonemes in each language have counterparts in the other two. Furthermore, of these eight, six are duplicated at the head of each list. Here, in addition to the universal phoneme tendency, one sees the reflection of a common proto-system. It is also possible by statistical means to establish certain hypotheses concerning chronological change.