Sanaston rikkaudesta ja sen mittaamisesta

TAUNO SÄRKKÄ

Kirjailija Eeva Joenpelto arvioi vuonna 1971 Parnassossa Pelkuruudesta-nimisessä artikkelissaan kriittisesti tuonaikaista kaunokirjallisuutta ja totesi mm. seuraavan (s. 67): »60-luku kantoi käsillään osallistuvan kirjallisuuden ohella minäkeskeistä uusmillerismiä. Näissä kummassakin kirjallisuuden lajissa on ylenmääräinen slangin käyttö kuivattanut kieltä. Sanavalikoima on paljon köyhempi kuin esimerkiksi 50-luvun parjatussa proosassa.» Joenpelto puhuu »vajaasta kielestä», jossa sellaiset mielialoja ilmaisevat sanat kuin *harmittaa*, *kiusata*, *väsyttää*, *hermostuttaa* jne. kuitataan alati yhdellä ainokaisella alatyylin verbillä. Jo varhemmin olin kuullut väitettävän Mika Waltarin kieltä sanastoltaan niukaksi, toisin sanoen hän olisi käyttänyt ns. frekvenssityyliä. Tämä taas olisi johtunut hänen helsinkiläisyydestään. On siten ilmeisestikin olemassa ongelma: kieli tai ainakin jotkin kielen tyylilajit saattavat köyhtyä sanastoltaan. Toisaalta voidaan tietenkin väittää, ettei kielen kokonaissanavarasto pienene. Sanasto vain vaihtuu kulttuurin muutoksen vuoksi. Niinpä esimerkiksi kadonneen pellavankäsittelytermistön tilalle on tullut uutta vaatetussanastoa. Tuollainenkin muutos koituu kielelle tappioksi siinä mielessä, että samalla menetetään lopullisesti kielen vanhaa perussanastoa. Kirjailija Joenpellon esittämä huoli on siis epäilemättä aiheellinen.

Tekstien sanavaraston mittaamiseksi on kehitelty koko joukko erilaisia indeksejä ja matemaattisia kaavoja, joilla pyritään saamaan selville sanaston »rikkaus» tai »köyhyys». Panen rikkaus ja köyhyys -sanat tuonnempanakin usein lainausmerkkeihin, koska ne eivät tässä yhteydessä ole ollenkaan yksiselitteisiä käsitteitä. — Tarkoituksenani on seuraavassa tutkia, kuinka hyvin eri indeksit soveltuvat suomenkielisten tekstien sanaston runsauden mittaamiseen ja minkätyyppistä »rikkautta» ne valaisevat. Yritän samalla selvittää myös, onko nykykirjailijoiden kieli tosiaan »köyhempää» kuin varhempien kirjoittajien ja onko tyylilajien kesken eroja sanaston »rikkaudessa».

Aineiston valinta on tämänkaltaisen tehtävän pahin pulma, eikä siitä johtuvaa virhettä voi millään keinoin kokonaan välttää. Tavoitteenani on ollut saada kultakin kirjoittajalta tai kustakin tekstistä yhdenmukaiset aineistot.

Otosten pohjaksi olen ottanut toisissa tapauksissa kirjoittajan koko tuotannon, toisissa taas jonkin laajan teoksen tai useita suppeita kirjoja.[1] Täysin samanveroisiin materiaaleihin ei ole aina ollut edes mahdollista päästä. Esimerkiksi Uusi testamentti on tuntuvasti suppeampi kuin monet muut teokset. Tulosten vertailun helpottamiseksi olen poiminut kirjoittajittain ja teksteittäin samankokoiset eli 5 000 saneen otokset. Otosyksikkönä on ollut vain kymmenen sanaa, sillä sanastontutkimuksessa tulokset ovat sitä luotettavampia, mitä pienempiä otosyksiköt ovat (otosyksikön koosta ks. esim. Pitkänen—Kohonen 1984: 28—29). Yksiköt on valittu satunnaislukujen taulukon avulla.

Materiaalini olen koonnut pitkän ajan kuluessa ja osaksi täysin toisentyyppisten seikkojen selvittämiseksi. Vaikka poikkeankin pääteemastani, kommentoin noita muita näkökohtia lyhyesti. Iskelmien sanaston keräsin nähdäkseni, missä määrin neli- ja viisikymmenluvun iskelmien sanasto on yhdenmukaista varhemman runokielen sanaston kanssa. Vertailusta kävi ilmi, että molemmissa aineistoissa ns. avainsanat ovat paljolti samoja. Niitä on iskelmissä käytetty vain huomattavan yksipuolisesti. (»Avainsanoista» ja niiden laskennasta ks. esim. Guiraud 1959: 90—92; Sigurd 1970: 131—132.) Puolueohjelmien ainesten avulla taas pyrin saamaan selville, voiko mekaanisella sanastojen vertailulla todeta eri puolueiden ohjelmien yhdenmukaisuuden tai erilaisuuden. Keskeisin tulos oli tietenkin se, että kaikki puolueet viljelevät samoja poliittisen kielen termejä, toisin sanoen puolueohjelmat laaditaan samankaltaisiksi. Yllättävintä oli ääripuolueiden sanastojen suurehko yhdenmukaisuus. Myös muutamien poliitikkojen kielestä olen hankkinut samanlaiset aineistot, joista ilmeni mm., että Kekkonen ja Vennamo käyttivät runsaasti samaa sanastoa ja että Sinisalon eittämätön »avainsana» oli *taistelu*. Vennamon »avainsanoja» olivat *kansa*, *yksilö* ja *valta* ja Kekkosen

[1] Luettelen aineslähteeni pelkistetysti taulukoissa esitetyssä järjestyksessä: Ilmari Kianto, Punainen viiva, Ryysyrannan Jooseppi; Kustaa Vilkuna, Lohi; Volter Kilpi, Alastalon salissa I—II; Väinö Kirstinä, Lakeus, Hitaat auringot, Puhetta, Luonnollinen tanssi, Talo maalla; Kaarlo Sarkia, Runot; Tuomo Polvinen, Teheranista Jaltaan, Jaltasta Pariisin rauhaan; Aaro Hellaakoski, Runot; Heikki Turunen, Simpauttaja, Joensuun Elli; Eeva-Liisa Manner, Tämä matka, Orfiset laulut, Kirjoitettu kivi, Fahrenheit 121, Jos suru savuaisi; F. E. Sillanpää, Hurskas kurjuus, Nuorena nukkunut; Väinö Linna, Täällä Pohjantähden alla I—II; Hannu Salama, Minä, Olli ja Orvokki, Siinä näkijä missä tekijä; Eeva Joenpelto, Vetää kaikista ovista, Kuin kekäle kädessä; Uuno Kailas, Runoja; Anni Polva, Anna mun kaikki kestää, Päävoittona mies; V. A. Koskenniemi, Kootut runot; Mika Waltari, Sinuhe egyptiläinen; Veijo Meri, Irralliset, Peiliin piirretty nainen; Puolueohjelmat = Suomen Sosialidemokraattisen Puolueen, Kansallisen Kokoomuksen, Keskustapuolueen, Suomen Kommunistisen Puolueen ja Suomen Maaseudun Puolueen viralliset puolueohjelmat; UT = Uusi testamentti; Iskelmät = Toivelauluja 2—6, 8—14, 16—25.

minä ja *Moskova*. – Edellisen kaltaiset selvitykset ovat luvattoman yksipuolisia, vaikka niitäkin näkee joskus harrastettavan (esim. Piironen 1984: 175–200).

Sanavaraston runsauden mittarit voidaan valtaosaltaan jakaa kahteen ryhmään. Niitä ovat 1) sanojen ja saneiden suhteeseen perustuvat kertoimet ja 2) hajontaan pohjautuvat indeksit.

Sanojen ja saneiden suhteelle rakentuvista luvuista yksinkertaisin on keskiarvo eli saneiden ja sanojen osamäärä, jota mm. G. U. Yule on käyttänyt ja jota hän on nimittänyt M-kertoimeksi (M = mean 'keskiarvo'; 1939: 363–384; 1968: 12). Tätä tunnetumpi ja käytetympi on TTR-kerroin (= type-token-ratio), joka taas on sanojen ja saneiden osamäärä eli edellisen käänteisluku. Gustav Herdan on käsitellyt sitä laajahkosti teoksessaan Type-token mathematics (sovelluksista suomen kieleen ks. esim. Särkilahti 1967: 258–259; 1969: 201–202). Kerroin on yksiselitteinen, mutta se on sidoksissa otoksen kokoon. Herdan on esittänyt samasta kertoimesta version, jossa käytetään sanojen ja saneiden määrän logaritmisia arvoja ja joka sen vuoksi on riippumattomampi otoksen koosta (1960: 20). Myös ranskalaisen Pierre Guiraud'n tarjoama R-indeksi (R = richesse 'rikkaus') perustuu sanojen ja saneiden suhteeseen (1959: 88–89). Kertoimen kaava on $\frac{V}{\sqrt{N}}$, jossa siis sanojen määrä jaetaan saneiden neliöjuurella. Guiraud'n väittämän mukaan se on riippumaton otoksen koosta, mikäli otos käsittää 10 000–30 000 sanetta. Jos otokset ovat pieniä, neliöjuuren otto ei näköjään pysty tasoittamaan otosten kokoeroja. Tällöin se tuottaa täysin TTR-kertoimen mukaisia tuloksia. Sen vuoksi ei R-kertoimen arvoja ole esitetty seuraavissa taulukoissa.

Hajontaan perustuvista indekseistä on ensiksi mainittava Yulen K-kerroin, jonka hän julkisti tietääkseni vuonna 1944 teoksessaan The statistical study of literary vocabylary (käytössäni on ollut teoksen myöhempi painos, 1968: 52–53 ym.; sovelluksista ks. esim. Bennet 1969: 29–41). Suoraan Yulen K-kertoimeen pohjautuu Herdanin V_m-kerroin (ranskalaiset tutkijat näkyvät nimittävänkin sitä Yulen ja Herdanin indeksiksi; Guiraud 1959: 89, Muller 1972: 202). Se on saanut kirjallisuudessa paljon huomiota osakseen (esim. Herdan 1956: 31–; 1964: 67–71, 162–164; 1966: 101–121; Guiraud 1959: 86; Muller 1967: 105–108; 1972: 202–203; Suomela 1975: 202–204; 1984: 28–30). Yleisessä muodossaan Herdanin kaava on seuraava: $V_m = \frac{s/\bar{x}}{\sqrt{n}}$ eli $\frac{s/\sqrt{n}}{\bar{x}}$ (sovellusten mukaan n = sanojen määrä). Kaava tarkoittaa periaatteessa samaa kuin keskiarvon keskivirheen ja keskiarvon osamäärä tai vaihtelukertoimen jakaminen n:n neliöjuurella. Indeksin etuna on yleisesti

pidetty sitä, että se on riippumaton otoksen koosta. Palaan tähän näkökohtaan vielä tuonnempana. Indeksi on ilmaisukyvyltään siinä suhteessa huono, että mitä pienempi on kertoimen arvo, sitä »rikkaampaa» sanasto on.

Edellä lueteltujen mittarien lisäksi on sanaston runsautta määritettäessä käytetty myös mm. entropian kaavaa. Virolainen Helle Niinemägi on puolestaan esitellyt Keel ja struktuur -sarjan neljännessä osassa (1970: 136—144) muutamia kertoimia, jotka selittävät osin sanaston määrää ja osin tekstin tyyliä. Eräässä niistä keskeisenä muuttujana ovat ns. hapaks legomenon -sanat eli tekstissä vain kerran esiintyvät sanat. Erityisesti Pierre Guiraud on pitänyt tällaisia sanoja tärkeinä sanaston rikkauden sekä tyylin vivahteikkuuden kuvastajina (1959: 89). Hän on nimittänyt niitä luonnehtimis- eli karakterisoimissanoiksi (= mots de caractérisation; näin varsinkin teoksessaan Les caractères statistiques du vocabulaire, joka ei tätä kirjoittaessani ole ollut käytettävissäni).

Aineistoni käsittelyssä niin kuin myös seuraavissa taulukoissa olen päähuomion kiinnittänyt edellä luettelemistani kertoimista kolmeen merkittävimpään, nimittäin TTR-arvoon, V_m-kertoimeen sekä hapaks legomenon -sanoihin (= HL). Taulukoissa on lisäksi ilmoitettu, mihin järjestykseen kirjoittajat tai tekstit ryhmittyvät eri kertoimien perusteella. Täydellisyyden vuoksi myös keskiarvot ja hajonnat on otettu mukaan. Korostettakoon, että taulukot muodostavat kiinteän kokonaisuuden.

TAULUKKO 1: muutamien kirjoittajien sanaston »rikkaus» (j = järjestys)

Teksti	Saneita	Sanoja	j	$TTR_{\%}$	$\bar{x}$
Kianto	5 000	2 244	1.	44,88	2,23
Vilkuna	5 000	2 108	2.	42,16	2,37
Kilpi	5 000	2 042	3.	40,48	2,45
Kirstinä	5 000	1 976	4.	39,52	2,53
Sarkia	5 000	1 965	5.	39,30	2,54
Polvinen	5 000	1 963	6.	39,26	2,55
Hellaakoski	5 000	1 913	7.	38,26	2,61
Turunen	5 000	1 894	8.	37,88	2,64
Manner	5 000	1 755	9.	35,10	2,85
Sillanpää	5 000	1 751	10.	35,02	2,86
Linna	5 000	1 747	11.	34,94	2,86
Salama	5 000	1 701	12.	34,02	2,94
Joenpelto	5 000	1 636	13.	32,72	3,06
Kailas	5 000	1 631	14.	32,62	3,07
Polva	5 000	1 550	15.	31,00	3,23
Koskenniemi	5 000	1 540	16.	30,80	3,25
Waltari	5 000	1 508	17.	30,16	3,32
Meri	5 000	1 449	18.	29,98	3,45
Puolueohjelmat	5 000	1 386	19.	27,72	3,61
UT	5 000	1 149	20.	22,98	4,35
Iskelmät	5 000	1 084	21.	21,86	4,61
$\bar{x}$	5 000	1 714		34,27	3,02

Ensimmäisen taulukon mukaan sanavarastoltaan »rikkainta» on Ilmari Kiannon teksti ja »köyhintä» iskelmien kieli. Näiden ääritapausten ero on huomattavan suuri. Sen sijaan taulukon keskustaan sijoittuvien kirjoittajien sanavalikoimien määrät ovat hyvin lähellä toisiaan (esim. Mannerin, Sillanpään, Linnan ja Salaman). Koska kaikilta kirjoittajilta tai kaikista teksteistä on koottu samankokoinen aineisto (5 000 sanetta), myös kaikki taulukossa esitetyt arvot (sanojen määrä, TTR% ja $\bar{x}$) ovat yhdenmukaiset. Kovin odottamattomia tulokset eivät ole. Kaunokirjailijoilla hajonta on ymmärrettävästi suuri, ja asiaproosan käyttäjät (Vilkuna ja Polvinen) tarvitsevat suurta sanavarastoa. Ehkäpä yllättävintä on se, ettei modernistirunoilijoiden (Kirstinän ja Mannerin) sanasto ole juuri »rikkaampaa» kuin perinnäisten runoilijoiden.

TAULUKKO 2: muutamien kirjoittajien sanaston »rikkaus» (j = järjestys, HL = hapaks legomenon -sanat)

Teksti	s	j	$V_{m\%}$	j	HL	HL%	j
Kianto	6,87	3.	6,51	3.	1 612	71,84	1.
Vilkuna	7,78	4.	7,14	5.	1 417	67,22	8.
Kilpi	9,82	9.	8,88	13.	1 458	71,40	3.
Kirstinä	8,78	6.	7,81	8.	1 334	67,51	6.
Sarkia	6,26	2.	5,55	2.	1 307	66,51	9.
Polvinen	6,02	1.	5,33	1.	1 214	61,84	15.
Hellaakoski	8,33	5.	7,29	7.	1 267	66,23	10.
Turunen	9,50	8.	8,27	10.	1 349	71,22	4.
Manner	10,61	14.	8,89	14.	1 125	64,10	13.
Sillanpää	10,44	12.	8,74	12.	1 188	67,85	5.
Linna	10,14	11.	8,48	11.	1 146	65,60	11.
Salama	12,00	15.	9,90	17.	1 221	71,78	2.
Joenpelto	12,19	16.	9,86	16.	1 102	67,36	7.
Kailas	9,97	10.	8,05	9.	994	60,94	16.
Polva	12,96	18.	10,21	20.	1 003	64,71	12.
Koskenniemi	9,24	7.	7,21	6.	919	59,68	18.
Waltari	12,94	17.	10,05	18.	910	60,34	17.
Meri	13,24	20.	10,08	19.	920	63,49	14.
Puolueohjelmat	13,12	19.	9,77	15.	776	55,99	20.
UT	16,98	21.	11,51	21.	664	57,79	19.
Iskelmät	10,48	13.	6,90	4.	532	49,08	21.
$\bar{x}$	10,37		8,40		1 117	64,40	

Toisen taulukon keskeisimmät kohdat ovat prosenttinen V_m-kerroin (= $V_{m\%}$) sekä hapaks legomenon -sanat (HL ja HL%). Taulukosta huomaa heti, että V_m-kertoimella mitattuna kirjoittajat asettuvat varsin selvästi sanaston »rikkauden» puolesta eri järjestykseen kuin TTR-lukujen mukaan. V_m -kertoimen perusteella rikkainta olisi Tuomo Polvisen asiaproosa. Sen sijaan esimerkiksi Volter Kilven kieli, jota on totuttu pitämään sanastoltaan vähintäänkin erikoisena ja joka TTR-luvun mukaan olisi kolmanneksi »rikkainta», osoittautuu V_m-kertoimen mukaan varsin keskinkertaiseksi. Yllättävintä

kuitenkin on, että iskelmien kieli on V_m-arvon perusteella peräti neljänneksi »rikkainta», vaikka edellisen taulukon lukujen mukaan se oli sanastoltaan selkeästi »köyhintä». Taulukosta näkyy, että V_m-kerroin osoittaa runokielen olevan yleensäkin suhteellisen »rikasta» Mannerin kieltä lukuun ottamatta.

Eri kertoimilla laskettujen arvojen kesken vallitsee melko jyrkkä ristiriita, joka kaipaa selvitystä. Lyhyesti sanottuna se johtuu siitä, että V_m-indeksi pohjautuu ennen muuta hajontaan, kun taas TTR-kerroin ei ota lainkaan hajontaa huomioon. Sellaisissa tapauksissa, joissa hajonta on keskiarvoon verrattuna suuri, myös V_m-kerroin saa suhteellisen suuren arvon, joka taas tarkoittaa sanavaraston »köyhyyttä». Normaalissa suomenkielisessä tekstissä hajonta pyrkii muodostumaan isoksi sellaisten suurtaajuisten sanojen vuoksi kuin *olla*-verbi sekä *ja*-konjuktio. Ne aiheuttavat jakauman vahvan positiivisen vinouden. Kuten jo varhemmin Charles Muller on huomauttanut (1967: 106—108; 1972: 203), puheena oleva kerroin on vahvasti sidoksissa kielen muutamiin suurtaajuisiin sanoihin, joita on kutsuttu grammaattisiksi sanoiksi ja muotosanoiksi (Nordberg 1968: 20). Jos tekstissä ei ole tuollaisia, frekvenssiltään muista kovasti erottuvia sanoja, V_m-luku jää pieneksi, mikä tarkoittaa sanaston »rikkautta». Niin on tavallisesti laita runokielessä, iskelmissäkin. Runokielen sanaluokkajakauma on muutenkin poikkeuksellinen. Omien tietokonetulostusteni mukaan esimerkiksi Kalevalan eri laitosten substantiivien määrä on n. 45 %. Sanaluokkajakauma näkyy olevan samantyyppinen yleensäkin runokielessä (Leskinen—Särkkä 1985: 54—56). — Todettakoon vielä, että Kalevalasta laskemani V_m-arvot osoittavat odotuksenmukaisesti Kalevalan kielen olevan sanastoltaan sangen »rikasta».

Jo edeltä on varmasti käynyt ilmi, että V_m-kerroin ei sovellu eri tyylilajien sanaston runsauden vertailuun, koska tyylilajeittain sanastojen jakaumat ovat erityyppisiä. Paremmin sanoen päädytään lopulta sellaiseen tulokseen, että V_m-kerroin on yhtä lailla tyylien erottelija kuin sanaston runsauden mittari. Samantapaisia näkökohtia on varhemmin esittänyt jo Ch. Muller (1967: 106—108).

Herdan ei havaitakseni ole juuri laskenut kertoimensa arvoja 5 000:ta sanetta pienemmistä aineistoista. Mullerin mukaan kerroin on riippumaton otoksen koosta, jos otos on suurempi kuin 10 000 sanetta. Sen sijaan 5 000:ta sanetta pienemmistä otoksista laskettaessa kerroin tuottaa hänen mukaansa pienempiä arvoja kuin suuremmista otoksista (1967: 106). Olen itse havainnut saman seikan; olen näet koonnut ensin kultakin kirjoittajalta viisi 1 000 saneen otosta, jotka olen sen jälkeen yhdistänyt. Voidaan siten epäilyksittä todeta, ettei tämäkään kerroin ole täysin riippumaton otoksen koosta. — Leo Suomela on pyrkinyt osoittamaan (1975: 202—204), että jo

1 000 saneen otokset riittäisivät kertoimen laskemiseen. En halua kokonaan kiistää Suomelan todistelua. Jos arvot lasketaan suhteellisen yhtenäisestä tyylilajista, otoksen koolla ei ole erikoisen suurta vaikutusta. Väitöskirjassaan Suomela on huomattavaa rohkeutta osoittaen soveltanut V_m-kerrointa kovin erityyppisiin muuttujiin ja erikokoisiin aineistoihin (1984: 30—).

Ns. hapaks legomenon -sanojen prosenttiosuutta kunkin kirjoittajan tai tekstin kokonaissanamäärästä (= $HL_{\%}$) pidän nimenomaan tyylintutkimuksen kannalta tärkeänä. Juuri niillä sanoilla saadaan aikaan tyylin vivahteikkuutta, joten ne ovat merkittävä osa sanaston »rikkautta». Taulukosta näkyy, että ne ovat ominaisia vallankin kaunokirjalliselle tyylille. Iskelmissä niitä on sekä absoluuttisesti että suhteellisesti vähiten.

Tarkasteluni pohjalta voi ensiksikin todeta, ettei ole riittäviä perusteita otaksua nykykirjailijoiden sanavalikoiman olevan merkittävästi »köyhempää» kuin varhempien kirjailijoiden. Eri mittarien mukaan mm. Heikki Turunen ja Hannu Salama käyttävät verraten »rikasta» sanavarastoa. Esimerkiksi Eeva Joenpellon sanasto olisi laskelmien mukaan jopa »köyhempää» kuin mainittujen kirjailijoiden.

Toiseksi voidaan pitää ilmeisenä, että käytetyn sanavaraston määrä on sidoksissa tyyliin. Mitään yllättävää ei havainnossa tietenkään ole. Ymmärrettävästikin Veijo Meren dialogia viljelevän kerronnan sanasto on yksipuolisempaa kuin esimerkiksi Kiannon ja Sillanpään maalailevan ilmaisun sanasto. On käynyt myös ilmi, että ainakin tarkasteltavana olleet asiapitoiset tekstit tarvitsevat keskimääräistä laajempaa sanavarastoa. Toisaalta taas puolueohjelmat samoin kuin iskelmätkin ovat erikoistyylilajeja, joissa on suorastaan pakko käyttää määräsanastoa, joka navaintojen mukaan on suhteellisen »köyhää».

Selvitykseni paljastaa kolmanneksi, etteivät esitetyn kaltaiset sanaston rikkausindeksit osoita välttämättä mitään teoksen tasokkuudesta. Erinomaisen esimerkin tästä tarjoaa Mika Waltarin Sinuhe egyptiläinen. Teoksen kielihän näyttäisi olevan kaikkien mittareiden mukaan perin »köyhää». Kuitenkin tiedämme, että teos on kieleltäänkin vaikuttava. Mahdollisesti kirjailija on pyrkinyt vaistomaisesti pöyhkeilemättömään, nöyrän vaatimattomaan sanavalikoimaan, jotta sanottava ja sanasto olisivat sopusoinnussa. Samankaltainen harmonian tuntu on myös Väinö Linnan tekstissä. Kirjailija ei herkuttele sanavaroilla asian kustannuksella.

Olen esitykseni lopussa päätynyt perin subjektiivisiin arviointeihin, vaikka varsinaisena teemanani on ollut sanaston runsauden mittaaminen eksaktien kaavojen avulla. Perin yllättävää se ei ole. Olen käsittelyssäni tullut siihen tulokseen, ettei olemassa olevien indeksien avulla pystytä ratkaisemaan lä-

heskään kaikkia sanaston runsauden ongelmia. Erilaiset mittarit tuottavat ristikkäisiä tuloksia, mittaavat erilaisia asioita, ja lisäksi on olemassa tuntuvia virhelähteitä. Toivon, että selvitykseni samalla myös varoittaa numerollisten menetelmien liiallisesta ihannoinnista. Kvantitatiivit menetelmät tarjoavat usein oivaa apua, mutta tuloksia on viime kädessä aina tarkasteltava tutkimusalan omista ehdoista lähtien.

LÄHTEET

BENNET, PAUL E. 1969: The statistical measurement of a stylistic trait in Julius Caesar and As you like it. — Ludomír Doležel and Richard W. Bailey (toim.), Statistics and style (Mathematical Linguistics and Automatic Language Processing 6) s. 29—41. New York.

GUIRAUD, PIERRE 1954: Les caractères statistiques du vocabulaire. Paris.

———— 1959: Problèmes et méthodes de la statistique linguistique. Dordrecht.

HERDAN, GUSTAV 1956: Language as choice and chance. Groningen.

———— 1960: Type-token mathematics. The Hague.

———— 1964: Quantitative linguistics. London.

———— 1966: The advanced theory of language as choice and chance. Berlin.

JOENPELTO, EEVA 1971: Pelkuruudesta. — Parnasso s. 66—68.

LESKINEN, HEIKKI—SÄRKKÄ, TAUNO 1985: Arhippa ja Miihkali Perttusen sampojakso. — Ongelmoita oppimia: näkökulmia Kalevalaan ja kansanrunouteen (Jyväskylä Studies in the Arts 23) s. 39—61. Jyväskylä.

MULLER, CHARLES 1967: Étude de statistique lexicale. Le vocabulaire du Théâtre de Pierre Corneille. Paris.

———— 1972: Einführung in die Sprachstatistik. Sammlung Akademie-Verlag 31. Berlin.

NIINEMÄGI, HELLE 1960: Statistilise stiilianalüüsi probleeme. — Keel ja struktuur 4 s. 136—144. Tartu.

NORDBERG, BENGT 1968: Recensenter och läsare: stil- och ordstudier i några lyrikrecensioner. — Ord och stil: Språkvårdssamfundets skrifter 1. Lund.

PIIRONEN, KAIJA 1984: Poliittisen sanaston muuttumisesta. — Historioitsija — taaksepäin katsova profeetta: Mauno Jokipiille omistettu juhlakirja (Studia Historica Jyväskyläensia 30) s. 175—200. Saarijärvi.

PITKÄNEN, ANTTI J. — KOHONEN, VILJO 1984: Johdatus kvantitatiiviseen kielentutkimukseen ja alan ATK-sovelluksiin. Hämeenlinna.

SIGURD, BENGT 1970: Språkstruktur. 3. p. Halmstad.

SUOMELA, LEO 1975: Herdanin tunnusluvun määräytymisestä. — Vir. 79 s. 202—204.

———— 1984: Dialogens förnyare. Stockholm Studies in Finnish Language and Literature 3. Stockholm.

SÄRKILAHTI, SIRKKA-LIISA 1969: Sanafrekvenssit kielen havainnollistajina. — Juhlakirja Paavo Siron täyttäessä 60 vuotta 2. 8. 1969 (Acta Universitatis Tamperensis A 26.) s. 200—204. Vammala.

———— 1967: Voidaanko kirjailijoiden kieltä tutkia kvantitatiivisin metodein? — Juhlakirja Kauko Kyyrön täyttäessä 60 vuotta 24. 11. 1967 (Acta Universitatis Tamperensis A 18) s. 253—266. Vammala.

YULE, G. UDNY 1939: On sentence-length as a statistical characteristic of style in prose. — Biometrica 30.

———— 1968 (alk. 1944): The statistical study of literary vocabulary.

Über den Reichtum des Wortschatzes und seine Messung

TAUNO SÄRKKÄ

Es existieren eine ganze Reihe diverse Indexe und Methoden, mit denen man die Reichhaltigkeit des Wortschatzes von Schriftstellern messen kann. Am bekanntesten dürften die TTR-Zahl (= Type-Token-Ratio), der Herdan V_m-Index sowie die Anzahl der sog. Hapax legomena sein. Die Untersuchung will klären, wie gut gerade die genannten Methoden geeignet sind, den lexikalischen Reichtum finnischsprachiger Texte zu messen. Gegenstand der Betrachtung ist ferner die Frage, ob die finnische Literatur nach Epochen oder Stilsorten Unterschiede im Umfang der Lexik aufweist.

Als Forschungsmaterial dienen Auszüge zu je 5000 Wörtern, sowohl von Schriftstellern als auch aus einigen Spezialtexten. Die Anzahl der Teilmaterialien beträgt insgesamt 21, so daß das Gesamtmaterial 105 000 Wörter umfaßt.

Die wichtigsten Ergebnisse der Betrachtung sind folgende: Der Umfang der Lexik wird durch die verschiedenen Indexe in recht unterschiedlicher Weise gemessen. So gehen die jeweiligen Ergebnisse auch in sehr unterschiedliche Richtungen. Vor allem der Herdan V_m-Koeffizient scheint unerwartete Resultate von finnischsprachigen Texten zu ergeben. Er wird zu stark nach einigen frequenzstarken Wörtern bestimmt. Er unterscheidet die Stilsorten und mißt auch den Reichtum des Wortschatzes. Es gibt auch keinen Index, der völlig unabhängig wäre vom Umfang des Korpus. Daß der Reichtum des verwendeten Wortschatzes epochenweise schwankt, läßt sich nicht mit völliger Sicherheit sagen, denn zumindest ein Teil auch der zeitgenössischen finnischen Schriftsteller verwendet eine relativ umfangreiche Lexik. Der lexikalische Reichtum ist dagegen eindeutig gebunden an die Stilsorten. Ein bemerkenswertes Ergebnis liegt darin, daß die Indexe über die Reichhaltigkeit des Wortschatzes nicht unbedingt etwas auszusagen vermögen über das künstlerische Niveau eines Werkes.