Keskustelua

Ajatuksia Suomen kielen taajuussanastosta

Tämän kirjoittaja on askarrellut sanastotilastoon liittyvien asioiden ääressä nelikymmenluvun loppupuolelta lähtien tutustuttuaan alan uranuurtajan Georg K. Zipfin teokseen »The Psycho-Biology of Language». Niinpä hankin viivyttelemättä Pauli Saukkosen ym. toimittaman »Suomen kielen taajuussanaston» (ks. Vir. 1979 s. 375—382) siinä toivossa, että sen avulla saattaisin viedä omat tutkimukseni päätökseen.

Päämääränäni on ollut sellaisen apuneuvon luominen, joka 1) ohjaisi vieraan kielen opiskelun oikeille raiteille, 2) antaisi mahdollisuuden kätevästi ja melko vaivattomasti verrata eri kirjoittajien sanastonkäyttöä, 3) tekisi mahdolliseksi eri kielten sanavaraston vertaamisen varsinkin ilmaisun tehokkuuden kannalta.

Kielen opiskelun osalta Taajuussanasto sanoo vain: »[tämän teoksen antamaa] tietoa voi soveltaa mm. oppikurssien laadinnassa, tekstien sanastollista helppoutta tai vaikeutta tarkistettaessa ja tekstien sisältöjä vertailtaessa». Mitään vihjettä ei anneta tämän toteuttamiseksi. Muita mainitsemiani käyttöjä ei edes mainita. Sen sijaan tosin annetaan tietoa eri sanaryhmien ja -luokkien esiintymistiheydestä eri tekstilajeissa ja eri aloilla: kaunokirjallisuudessa, radiossa, lehdistössä ja tietokirjallisuudessa. Nämä sinänsä arvokkaat tiedot ovat kuitenkin siinä mielessä lopullisia, että ne eivät viittaa mihinkään näiden perusteella mahdollisesti kehitettäviin tutkimustehtäviin.

Sanastotilaston sovelluksesta vieraan kielen opiskeluun olen todennut, että mikäli sananvalinnassa tai itseopiskelussa on saavutettu tilanne, jossa enintään 10 % tekstin sanoista on opiskelijalle tuntematomia, sanakirjaan ei yleensä tarvitse turvautua, koska tuntemattomat sanat voi arvata tekstiyhteydestä. (Ks. kirj., Kalevalaseuran vuosikirja 1974 s. 370—379.) Tällainen kielitaito vaatii suomen kielessä runsaan 1 000 sanan osaamista, mikäli ei ole kysymys erityisesti opiskelijaa varten laaditusta tekstistä. Kun minun tutkimukseni tältä kohdin ovat täysin valmiit vain Uuden testamentin kielen osalta enkä ole saanut tilaisuutta saattaa tutkimustani loppuun, kiirehdin tarkistamaan, voisiko »Suomen kielen taajuussanasto» auttaa työn perille viemistä. Ja Taajuussanasto sanoi: 10 000 sanaa kattaa 89,4 % tekstistä eli jättää 10,6 % tuntemattomaksi. Kun itse olin moneen kertaan eri tekstejä laskien voinut todeta, että vaikeimmissakin kielissä 10 %:n vajaus saavutetaan huomattavasti pienemmällä sanaluvulla (englannissa alle 3 000), niin välähti mieleeni teos »Kuinka tilastolla valehdellaan». Ja niin oli mielestäni tärkeää selvittää, mitä periaatteellisia virheitä tässä voi piillä.

Ehkä ei varsinainen virhe vaan sanaston käyttöä vaikeuttava tekijä on järjestysnumeron maaginen ihannointi. Sanan useuden tarkistus aakkosluettelosta on näin ollen etsinnän tai ainakin yhteen-

laskun takana. Toiselta puolen järjestysnumero on liian satunnainen. Taajuussanastossa *ja* on toisella sijalla; kuitenkaan se ei esiinny ensinkään Kalevalassa. Kalevalan kielihän tosin poikkeaa yleiskielestä. Mutta jos kiinnitämme huomiomme sanojen esiintymiseen edellä mainituilla neljällä alalla, niin toteamme, että sana nro 5 *joka* saa sijat 10, 7, 4 ja 4, sana 7 *tämä* sijat 27, 5, 8 ja 8, sana 17 *myös* sijat 107, 34, 13 ja 14 jne. Yksittäisten sanojen useus siis vaihtelee oireellisesti. Tästä pulmasta selviämme kuitenkin vähällä: perustamme tutkimuksemme sopivan kokoisiin sanaryhmiin. Erot supistuvat täten niin, että ne eivät vaikuta tekstin arviointituloksiin.

Kirjan johdannossa viitataan Sture Allénin ruotsin kielen taajuusluetteloihin. Olisikohan Allénin vaikutusta, että arvosteltavamme teos ei ole ensinkään sovellettu suomen kielen rakenteen ominaispiirteisiin? Eroaahan kielemme olennaisesti enimmin tuntemiemme kielten rakenteesta ainakin kahdella tilastollisesti vaikuttavalla tavalla: runsaiden taivutusmuotojen ja yhdyssanojen takia. Viittaamassani Kalevalaseuran vuosikirjan kirjoitelmassa olen osoittanut, miten erään yli 22 000-jäsenisen ilmaisuryhmän muodostamiseen englanti tarvitsee noin 130 morfeemia, mutta suomi tulee toimeen 64:llä. Jatkuvan tekstin osalta totesin, että Juho Hollo on tarvinnut Munthen »The Story of San Michele» (Huvila meren rannalla) -kirjan satunnaisesti valitun 1 000 sanan pituisen otteen suomentamiseen vain n. 600 sanaa. Tällöin ei taivutusmuotoja ole laskettu eri sanoiksi, ja yhdyssanojen yhdysosat on luettu erilleen. Ja suomen yhdyssanoissa on jälkiosa aina pääsana (päinvastoin kuin esim. ranskassa paitsi yhdyssanoissa myös sanaliitoissa). Taajuussanastosta olen löytänyt sanan *laki* aakkosluettelon 20 eri paikasta. Perusmuodossaan sillä on numero 198 ja useus 252. Mutta tottahan *eläkelaki*, *perustuslaki*, *rikoslaki*, *verolaki* jne. ovat lakeja nekin, ja kun otamme nämä huomioon, yhteisuseus on (ainakin) 462. Jos koko luettelo jätettäisiin muuten ennalleen, niin *laki* saisi sijan 89. Vielä mahtavampi olisi *kunta*-sanan siirtymä: useus 242 on antanut sijan 210, mutta koko ryhmän useus on lähes 1 000 ja sija olisi 19:n ja 20:n välissä. Tottahan ainakin opiskelijan kannalta *valtakunta*, *maakunta*, *henkilökunta*, *ihmiskunta*, *luomakunta*, *lautakunta* ja moni muu *-kunta* kuuluvat yhteen, vaikka *kymmenkunta* tuntuisikin eroavalta. Onhan mieletöntä, että koko aakkosluettelo on tutkittava läpeensä kaikkien *-kuntien* löytämiseksi.

Näiden virheiden korjaaminen, ts. koko aakkossanaston uudistaminen, on välttämätöntä, jos sanastoa tahdotaan käyttää alussa mainitsemieni päämäärien saavuttamiseksi.

Vilho Setälä