Tekstisyntaksin tilastoanalyysiä

AULI HAKULINEN — FRED KARLSSON — MARIA VILKUNA *Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus.* Publications of the Department of General Linguistics, University of Helsinki No. 6. Helsinki 1980. 189 s.

Kieli soveltuu hyvin tilastollisen tutkimuksen kohteeksi: sekä puhuttua että kirjoitettua tekstiä on runsaasti saatavilla, tekstit on helppo osittaa kielellisesti merkittäviksi yksiköiksi, eikä erilaisten yksikköjen esiintymistaajuuden tai niiden välisten riippuvuussuhteiden laskeminen tuota suuriakaan vaikeuksia. Ongelmallisinta onkin ollut laskettavien piirteiden valinta ja saatujen lukujen tulkinta: mitä niiden avulla voidaan paljastaa itse tutkimuskohteesta, kielestä?

Suomen kielen tilastollinen tarkastelu ei ole ollut erityisen runsasta; silti siinä on nähtävissä kolme pääsuuntausta. Yksi niistä on erilaisten luettavuusmittarien kehittely. Tavoitteet ovat tällöin etupäässä käytännöllisiä: pyritään rakentamaan mahdollisimman yksinkertainen tekstin kvantitatiivisiin ominaisuuksiin nojaava laskentamalli, joka mahdollisimman luotettavasti pystyy osoittamaan, onko teksti nopea- vai hidaslukuista. Tyypillisiä tarkastelun kohteeksi joutuvia tekstin piirteitä ovat esimerkiksi virkkeiden, lauseiden ja sanojen pituus, verbien ja adjektiivien suhteellinen osuus jne. Monetkaan näin kehitellyistä mittareista eivät ole lingvistisesti erityisen antoisia.

Toinen suuntaus käsittelee sekin yleensä pelkästään frekvenssijakaumia. Sen tärkein saavutus ovat erilaiset taajuussanastot. Tutkimuksen kohteena saattaa olla myös tekstin sanojen tai virkkeiden pituusjakauma, samoin esimerkiksi foneemien tai grafeemien ja morfologisten paradigmojen yksiköiden (sijapäätteiden, aikamuotojen ym.) esiintymistaajuus; periaatteessa mistä tahansa fonologisen ja grammaattisen struktuurin yksiköistä voidaan tällä tavoin laatia kvantitatiivinen kuvaus. Näin voidaan selvittää mm. erilaisten yksikköjen funktionaalista kuormitusta, redundanssin määrää kielen tietyssä osasysteemissä jne.; ylipäänsä tällä tavoin on saatavissa tärkeää perustietoa, jolla on usein käyttöä muiden lingvistisesti merkittävien tutkimusongelmien ratkaisemisessa.

Kolmas tarkastelutapa kytkeytyy läheisesti edelliseen, mutta siinä on lisäaspektina mukana vertailu. Vertailu on säännöllisesti tekstien välistä: halutaan selvittää, miten tiettyjen kielen yksikköjen esiintymistaajuus vaihtelee erilaisissa teksteissä. Taustamuuttujina ovat usein tekstin syntyaika tai -tilanne, samoin puhuja tai kirjoittaja ja hänen taustansa: ikä, sukupuoli, sosiaalinen asema, ammatti, koulutus, kasvuympäristö jne. Tavoitteena on löytää sellaisia tekijöitä, jotka ovat sidoksissa vertailtavissa teksteissä näkyviin kielellisiin eroihin. Tyylintutkimus on jo pitkään käyttänyt tällaista tutkimusotetta, samaten sosio- ja psykolingvistiikka. Juuri kielellisen variaation ja siihen liittyvästi myös kielen muutosprosessien tutkimus joutuu tukeutumaan kvantitatiiviseen analyysiin.

Lähtökohdiltaan »Suomen tekstilauseiden piirteitä: kvantitatiivinen tutkimus»

kuuluu näistä kolmesta suuntauksesta keskimmäiseen, vaikka päähuomio kiinnitetäänkin siinä yksityisten muuttujien asemesta ensi sijassa eri muuttujien välisiin riippuvuussuhteisiin. Jossain määrin pidetään tosin silmällä myös tekstityyppien välisiä eroja, mutta tutkimuskohteeksi on kuitenkin valittu varsin homogeeninen korpus: neutraali kirjoitettu asiaproosa.

Työ on saanut alkunsa vuonna 1975 Suomen Akatemian rahoittamassa tekstilingvistiikan tutkimusryhmässä; suunnitelma oli alkuaan laajempi, sillä tavoitteena oli verrata suomen ja ruotsin — joiltakin osin myös englannin — keskeisimpiä tekstisyntaktisia ja -semanttisia piirteitä. Kontrastiivinen näkökulma jäi kuitenkin sivuun, ja suomen kieltä koskeva suunnitelman osa eriytyi omaksi »Suomen kielen kvantitatiivista tekstisyntaksia» -nimiseksi hankkeekseen. Kysymys on tämän hankkeen loppuraportista.

Tutkimuksen ensisijaiset tavoitteet ovat tekstilingvistisiä; erityisen vahvasti siinä pannaan painoa eräisiin sanajärjestysilmiöihin: »Millä tavoin nominaalisen lausekkeen sijaintiin lauseessa vaikuttaa sen tarkoitteen tunnettuus, sen tekstuaalinen sidonnaisuus ja sen lauseenjäsenyys? Mikä näistä "voittaa" tai on ratkaisevin? Mitkä seikat vaikuttavat eniten esim. objektin topikaalistamiseen? Onko suomessa päälauseilmiöitä, joita sivulauseista ei lainkaan tavata? Miten konstituentin kompleksisuus ja pituus puolestaan vaikuttavat sen paikkaan lauseessa?»

Tutkimuksen ansiot eivät suinkaan ole siinä, että se vastaisi näihin kysymyksiin; itse asiassa ne jäävät kaikki edelleen avoimiksi. Suurimmalta osalta tämä johtuu tutkittavan ilmiön luonteesta: sanajärjestykseen vaikuttavat tekijät ovat siksi mutkikkaita, ettei niitä ole helppo tavoittaa pelkästään tilastollisen tarkastelun ja suhteellisen pienen muuttujajoukon avulla. Kooditusjärjestelmä kokoaa kuitenkin tietoa lukuisista muistakin suomen kielen perussyntaksiin kuuluvista lauseiden ominaisuuksista, ja työn merkitystä voidaankin arvioida sen perusteella, mitä se tuo esiin tarkattavina olevista kielenpiirteistä ja miten se osaltaan edistää suomen kielen kvantitatiivisen tutkimuksen kehitystä.

Aineistokseen tutkimusryhmä on siis ottanut kirjoitetun asiaproosan. Sitä edustamaan on poimittu 123 eri tekstiä, yhteispituudeltaan 10 149 lausetta (5 016 virkettä, 66 851 sanaa). Tekstit ryhmittyvät kuuteen alakategoriaan, joiden väliset erot osoittautuvat verraten lieviksi. Korpus on tarkoituksenmukaisesti ja ilmeisen onnistuneesti valittu.

Tutkimuksen perusta ja samalla myös sen tärkein saavutus on itse kooditusjärjestelmä. Siinä on kaikkiaan 63 muuttujaa; niistä neljä on puhtaasti teknisiä, muut ryhmittyvät kuuteen luokkaan. Ne kuvaavat 1) virkkeiden, 2) lauseiden ja 3) konstituenttien rakennetta, 4) lause- ja sanamäärää, 5) nominaalisten konstituenttien tekstuaalisia ominaisuuksia ja 6) transformaatioita.

Kooditusjärjestelmä rakentuu tietenkin sanajärjestyksen ympärille, sillä tutkimuksen perustavoitteenahan on juuri sanajärjestyksen ja siihen vaikuttavien tekijöiden osoittaminen. Neutraaliksi järjestykseksi on koodituksessa otettu SV(X), ja poikkeamat tästä on pyritty selvittämään. Selitettävä muuttuja on siten sanajärjestys: kaikki kuusi SVX-yhdistelmän permutaatiota ja yksinkertaisemmat yhdistelmät V sekä SV ja VX permutaatioineen, yhteensä siis 11 eri varianttia.

Selittävistä muuttujista selvästi tekstuaalisia ovat nominaalisten konstituenttien sidoskeinoja sekä niiden referentin tunnettuutta ja mainittuutta osoittavat; syntaktisista muuttujista sanajärjestystä pyrkivät selittämään mm. konstituenttien rakennetta ja sanamäärää sekä lausetyyppiä, -muotoa ja -rakennetta kuvaavat.

Parhaiten kooditusjärjestelmä pystyy nähdäkseni tavoittamaan lauseen subjektin, objektin ja predikatiivin. Niistä kustakin on kooditettu erikseen kyseisen nominaalilausekkeen rakenne, sanamäärä, sijamuoto, referentin tunnettuus ja mainittuus sekä tekstuaalinen sidoskeino ja kyseisen konstituentin semanttinen status. Oma mielenkiintoinen ryhmänsä ovat transformaatiomuuttujat (topikaalistus, sen todennäköinen syy, subjektin lykkäys, verbin siirto, dislokaatio ja ellipsi); niiden avulla on saatavissa ainakin jonkin verran tietoa lauseiden tekstuaaliselta kannalta

olennaisista piirteistä. Huomattavasti heikommaksi jää adverbiaalien käsittely nimenomaan siksi, että tutkimusryhmä on ottanut kustakin lauseesta tarkasteltavakseen vain sen ensimmäisen ja viimeisen adverbiaalin. Muut adverbiaalit jäävät siis kokonaan tämän kooditusjärjestelmän tavoittamattomiin. Ratkaisu tuntuu sikälikin yllättävältä, että käsittelyn perusyksikkönä on lause eikä sitä hallitseva laajempi syntaktinen konstruktio, virke. Voisi nimittäin odottaa, että lause pyrittäisiin tällöin kuvaamaan mahdollisimman täydellisesti. Menettelylle on tietenkin olemassa ilmeiset perusteet: tekijät pyrkivät selvittämään nimenomaan lauseiden tekstuaalisia ominaisuuksia; kun tutkimuksen ekonomian kannalta on tingittävä joissakin kohdin, tämä tapahtuu juuri syntaktisen rakenteen kuvauksessa. Sitä paitsi voidaan hyvin olettaa, että adverbiaaleista ovat tekstuaaliselta kannalta tärkeimpiä juuri lauseen ensimmäinen ja viimeinen.

Jotkut muuttujista ovat osoittautuneet tutkimustavoitteiden kannalta merkityksettömiksi; ainakaan tekstifunktiota, lauseen geneerisyyttä ja pragmaattisia elementtejä käsittelevät eivät näytä olleen erityisen antoisia. Samaten nominaalisten konstituenttien semanttista statusta (abstraktiotasoa) kuvaavat neljä muuttujaa ovat jääneet analyysistä sivuun (jollain tämäntapaisella muuttujalla saattaa hyvinkin olla käyttöä ainakin stilistisessä tutkimuksessa).

Koodituksen kannalta muuttujista osa on ongelmattomia: on esimerkiksi helppo ratkaista, milloin subjekti on nominatiivi-, partitiivi- tai genetiivimuotoinen. Aina alakategorioiden rajat eivät kuitenkaan ole yhtä selvät, ja tämä tietenkin vaikeuttaa kooditusta ja heikentää samalla tulosten merkitsevyyttä. Esimerkiksi tällaisesta hankalasti käsiteltävästä muuttujasta sopii adverbiaalin tyyppiä kuvaava: kyseessä voi olla valenssi-, kehys- tai kommenttiadverbiaali, konnektiivi, pakollinen tai irrallinen predikatiiviadverbiaali tai jokin näiden kuuden ryhmän ulkopuolelle jäävä adverbiaalityyppi. Rajankäynti näiden luokkien välillä ei ole helppoa, eikä ehdottomaan täsmällisyyteen tietenkään päästä. Niin kuin tekijät toteavatkin, »valenssiadverbiaalien erottaminen vapaista määritteistä on usein makukysymys». (Tältäkin kannalta hieman yllättävää on se pohdiskelu [s. 144], jota osin varsin spekulatiivisten frekvenssilukujen nojalla esitetään nominaalijäsenten primaarisuushierarkiasta ja valenssiadverbiaalin asemasta siinä.)

Kooditettavan yksikön valinta ei tunnu aivan kiistattomalta. Ilmeisesti halu selvittää pää- ja sivulauseiden eroja on johtanut siihen, että perusyksiköksi on otettu lause eikä esimerkiksi virkettä tai päälausetta siihen mahdollisesti kytkeytyvine sivulauseineen. Nyt tämän syntaktisen konstruktion rakenne jää osin avoimeksi; vain sen sisältämien lauseiden luku, sanamäärä ja tyyppi (*että*-lause, epäsuora kysymyslause jne.) saadaan kuvatuksi. Lisätietojen saamiseen tarvitaan erilaisia ristiintaulukointeja. Kooditusjärjestelmä on ylipäänsäkin varsin raskaskäyttöinen juuri siksi, että monet yksinkertaisetkin seikat saadaan esiin vain hankalien ristiintaulukointien avulla.

Yllättävänä voi pitää sitäkin, että ns. lauseenvastikkeista on lauseiksi katsottu objektina oleva partisiippirakenne, temporaalirakenne, aikaa ilmaisemaan käytetty modaalirakenne sekä subjektina oleva 1. infinitiivi määritteineen. Ratkaisu on osittain tekninen ja nimenomaan tämän tutkimuksen tarpeista määräytyvä. Tärkein peruste on ollut se, että tällaiset rakenteet »sisältävät paljon informaatiota ja paljon uusia tai mainittuja referenttejä»; jos rakenteet olisi merkitty tavallisiksi adverbiaaleiksi, näitä seikkoja ei olisi saatu näkyviin. Itse ongelma ja sen ratkaisu antavat lukijalle aavistuksen tutkimusryhmän kohtaamista vaikeuksista: tarkasteltavina olevat tekstuaaliset ilmiöt kytkeytyvät niin suureen joukkoon erilaisia kielellisiä tekijöitä, ettei niitä kaikkia voida mahduttaa analyysiin. On tyydyttävä kompromisseihin, tilapäisratkaisuihin, jotka usein ovat pakostakin epäsystemaattisia ja vähentävät kooditusjärjestelmän soveltuvuutta muihin tutkimustehtäviin.

Varsin vähäisiksi osoittautuvat pää- ja sivulauseen sanajärjestyserot, joiden selvittäminen on ollut päämääränä koodituksen perusyksikköä valittaessa. Oikeastaan käy ilmi vain se, että verbiloppuisuus on

päälauseenetisessä ja -sisäisessä sivulauseessa suunnilleen kaksi kertaa yleisempää kuin päälauseessa: »Tällaisena heijastuu siis nykykielen kirjoitettuun muotoon sanajärjestystyyppi, jonka on oletettu olleen suomen kielen esivaiheissa vallitsevana: jäljellä on lievä tilastollinen suuntaus.» Kiinnittämällä huomiota verbin nominaalirakenteisiin olisi voinut löytää vahvempiakin jälkiä verbiloppuisuudesta; nämä ovat kuitenkin pääosiltaan jääneet tarkastelun ulkopuolelle.

Teoksen antoisin osa on itse muuttujien erittely, joka viekin lähes puolet käytetystä sivumäärästä. Tutkimusryhmä perustelee siinä ratkaisujaan, valaisee runsaalla esimerkkiaineistolla itse kooditusta, arvioi eri muuttujien kuvausvoimaa ja esittää parannusehdotuksia. Lukijalle työn tämän osan merkitys korostuu jo senkin vuoksi, että tekijät eivät juuri pohdi muualla tutkimuksensa heikkouksia ja saavutuksia: varsinainen kokoava loppukatsaus kirjasta puuttuu.

Itse tulokset on esitelty lukuisina taulukoina, ja tarkastelu on yksinomaan kvantitatiivista: konkreettinen kieliaines on kokonaan jätetty syrjään eikä numeroista piirtyvää kuvaa väritetä ainoallakaan tekstiesimerkillä. Riittävän otoksen kokoa ja alatekstilajien eroja käsittelevä kolmas luku ei osoittaudu erityisen antoisaksi. Siinä tosin vahvistuu havainnolliseksi se tosiseikka, joka on ennustettavissa tilastotieteen yleisten periaatteiden nojalla: yksinkertaisen frekvenssijakauman selvittämiseen ei tarvita kovinkaan laajaa korpusta. Usein jo muutaman sadan lauseen aineisto antaa riittävän tarkkuuden, sillä frekvenssien »jäätymispiste» on melko alhainen. Vasta muuttujien ristiintaulukointi vaatii korpuksen tuntuvaa kasvattamista. Kuusi alatekstilajia (tietokirjat, ensyklopediat, pääkirjoitukset, kulttuuriartikkelit, pakinat ja tiedottavat artikkelit) eivät ratkaisevasti eroa toisistaan. Käy kuitenkin ilmi, että muuttujista monet ovat varsin käyttökelpoisia tyyli-indikaattoreiksi.

Päätösluvussa esitellään yli 70 taulukon avulla eri muuttujien frekvenssijakaumia ja muuttujien välisiä riippuvuussuhteita. Varsinaiset tutkimusongelmat eivät kvantitatiivisen analyysin avulla ratkea, joskin eräitä vahvuudeltaan eriasteisia tendenssejä paljastuu. Taulukoilla on kuitenkin pysyvämpääkin käyttöarvoa. Ne täyttävät ensinnäkin tutkimusryhmän odotukset siinä suhteessa, että niistä on löydettävissä runsaasti tietoa monien suomen kielen syntaktisten ilmiöiden esiintymistaajuudesta. Esimerkiksi lauseet, joista pintasubjekti puuttuu, ovat yllättävän yleisiä; niin kun tekijät toteavat, on vahvasti liioiteltua katsoa — kuten on usein tehty — suomen kielen kuvauksessa subjekti »lauseen toiseksi pakolliseksi pääjäseneksi». Hyvin selvänä on nähtävissä myös viskurilain vaikutus: raskaat konstituentit pyrkivät sijoittumaan lauseen loppuun. Joissakin kohdin halu antaa taajuushavainnoille »lingvistinen selitys» tuntuu tosin johtavan liiallisuuksiin. Epäselväksi jää esimerkiksi se, miten »tunnusmerkkisyysteorian näkökulmasta saattaisi hyvinkin ennustaa», että yksilauseisten virkkeiden osuuden pitäisi olla suurempi kuin todetut 38 %.

Taulukot antavat arvokasta taustamateriaalia myös variaation tutkijalle. Niiden voi katsoa edustavan neutraalia asiaproosaa, ja nimenomaan eri tekstilajeihin keskittyvä tarkastelu, »tyylintutkimus», voi käyttää hyväkseen taulukoihin tiivistettyjä frekvenssitietoja. Valitettavaa tältä kannalta on tosin se, etteivät kaikki taulukot ole täysin viimeisteltyjä: niissä on puutteita sekä tekniikassa että tarkkuudessa. Rivi- ja sarakesummat puuttuvat aivan yleisesti, ja usein esitetään tiedot vain osasta muuttujan alakategorioita (esim. adverbiaalien lukua lauseessa osoittavasta taulukosta 55 puuttuu tieto niiden lauseiden määrästä, joissa ei ole adverbiaalia lainkaan). Hioutuneempi tekniikka olisi tältä osin ollut tärkeää nimenomaan taulukoiden kontrolloitavuuden kannalta. Kun samaa muuttujaa käsitellään eri taulukoissa, loppusummien tulisi täsmätä, mikäli taulukointi on tehty rikkeettömästi. Nyt esimerkiksi lausekesubjektin konstituenttirakennetta ja sanamäärää kuvaavat taulukot 26 ja 51 eivät täsmää sen enempää keskenään kuin subjektin konstituenttirakennetta osoittavan taulukon 25 kanssa, eivätkä epätarkkuudet suinkaan

ole aivan vähäisiä.

Puutteita tutkimuksessa siis on, mutta osaksi ne selittyvät nimenomaan sen tekstilingvistisistä tavoitteista: ei voida kohtuullisesti odottaa, että kooditusjärjestelmä antaisi edes lähimain tyydyttävän kuvan lauseiden syntaktisista ominaisuuksista. Monet niistä on jätetty taka-alalle; esimerkiksi verbin nominaalimuotojen asema lauseessa saadaan esiin vain hyvin puutteellisesti. Tältä osin kritiikkiä on helppo esittää. Se on kuitenkin epäoikeudenmukaista, koska tutkimus ei ole pyrkinytkään puuttumaan tällaisiin kysymyksiin. Vaikka järjestelmä tuskin sellaisenaan soveltuu sen enempää tulevan syntaktisen kuin tekstilingvistisenkään tutkimuksen pohjaksi, sillä on kuitenkin merkitystä tällaisen tutkimuksen suunnittelussa. Nimenomaan se, että kooditussysteemi on testattu riittävän laajalla korpuksella, on tuonut ilmi lukuisia tekstisyntaksin kvantitatiiviseen tarkasteluun liittyviä ongelmia. Osa niistä on ratkaistu, ja myös avoimeksi jääneistä on vastaisuudessa helpompi selvitä.

Kvantitatiivisen tekstisyntaksin tutkimushanke on ollut pioneerityö. Siinä kehitellyissä muuttujissa on runsaasti sellaisia, joita ei tähän mennessä ole käytetty ATK-menetelmiä soveltavassa syntaktisessa tutkimuksessa. Niin kuin tekijät toteavat, »kvantitatiivisen analyysin pätevyys riippuu ratkaisevasti siitä teoreettisesta analyysista, jolle tilastointi perustuu». Juuri tämä yritysten ja erehdysten myötä hioutunut teoreettinen pohja ja sille rakentuva kooditusjärjestelmä on edistysaskel suomen kielen kvantitatiivisen tutkimuksen ylen kivikkoiseksi osoittautuneella tiellä siitäkin huolimatta, että itse tutkimustyö on jäänyt puolitiehen. Loppuraportista nimittäin heijastuu varsin selvästi eräänlainen väsymys: läheskään kaikkia potentiaalisesti kiintoisia ristiintaulukointeja ei ole tehty, esiin tulleita yksityisiä ongelmia ei ole lähdetty selvittelemään, yleiskuvaa tutkittavista ilmiöistä ei ole jaksettu piirtää eikä konkreettiseen tekstimateriaaliin ole enää tilastoinnin jälkeen palattu.

Nyt valmistunut työ avaakin jo sinänsä suoran jatkomahdollisuuden. Siinä käytetystä tutkimusaineistosta on tietokoneeseen siirretty vain lauseiden kooditetut ominaisuudet, ei itse tekstejä. Olisi varsin vaivaton ja yksinkertainen asia lisätä sinne myös vastaavat tekstit, jolloin kooditusjärjestelmä voisi toimia hakusysteeminä: haluttaessa saataisiin nopeasti esiin vaikkapa kaikki sellaiset lauseet, joissa partitiivisubjekti esiintyy verbin edellä tai joissa lauseobjekti on topikaalistettu. Toisin sanoen nyt laadittua kvantitatiivista analyysiä voitaisiin täydentää kvalitatiivisella, ja itse korpus toimisi tavallaan lauseopin ja tekstilingvistiikan arkistona, josta vähällä vaivalla olisi löydettävissä runsas aineisto monien sellaisten havaintojen yksityiskohtaisempaan selvittelyyn, jotka nyt tuntuvat jäävän irrallisiksi.

Kvantitatiivisen tutkimuksen merkitys ei olekaan pelkästään siinä, että sen frekvenssijakaumat paljastavat ilmiöiden välisiä riippuvuuksia ja niitä ohjaavia tendenssejä. Se tuo nimittäin näkyviin myös poikkeamia, muuttujien epätavallisia yhdistelmiä, joiden selvittäminen johtaa pois numeerisesta analyysistä yksityisten esiintymien perinpohjaiseen tarkasteluun. Niihin syventyminen jää kuitenkin tavoitteeltaan suppearajaisempien erityistutkimusten tehtäväksi.

Pentti Leino