LÄHTEET

Tekstilajit tilastolukuina

DOUGLAS BIBER *Variation across speech and writing*. Cambridge University Press, Cambridge 1988. 299 s.

Tietokoneiden ansiosta kielentutkimuksessa voidaan käsitellä suuriakin tekstimääriä ja kielen kuvaukset voidaan perustaa olennaisesti suurempiin aineistoihin kuin aikaisemmin. Tuttua on sanakirjojen ja morfologisten, jopa syntaktisten kuvausten laatiminen miljoonien ja miljoonien saneiden korpuksista. Uudehkoa näkökulmaa tuovat sitä vastoin t e k s t i e n kvantitatiiviset kuvaukset. Mitä annettavaa on massiivisilla aineistoilla ja tilastolaskelmilla nimenomaan tekstien kuvauksessa?

Suuret aineistot mahdollistavat typologisten tendenssien kuvauksen. Tämä puolestaan antaa mahdollisuudet vertailla erikielisiä tekstejä tai kuvata samankielisten tekstien variaatiota. Erikielisten tekstien typologisten piirteiden kuvauksesta on hyvä esimerkki uudehko tutkimus, jossa verrataan erinäisten rakennepiirteiden avulla espanjan- ja englanninkielisiä tekstejä (Myhill 1992). Metodiltaan toisentyyppinen on Biberin tutkimus, jonka tavoitteena on kuvata englanninkielisten tekstien variaatiota.

Puhutun ja kirjoitetun kielen erojen kuvaus tai pikemminkin koherentin kuvauksen puuttuminen sai Douglas Biberin ryhtymään laajaan tekstilajien tilastolliseen analyysiin. Toinen puute aikaisemmissa tutkimuksissa Biberin mielestä oli se, että sekä kirjoitettua että puhuttua kieltä oli kumpaakin käsitelty yhtenä kokonaisuutena erittelemättä kielen lajeja. Biber halusi ottaa tutkimuksessaan huomioon eri tekstilajit; toiseksi hän halusi tarkastella tekstejä nimenomaan kvantitatiivisin menetelmin, ei pelkästään kvalitatiivisesti eroja toteamalla ja kuvaamalla. Tutkimukseen tarvittiin suuri aineisto sekä sellainen ongelmanasettelu, että aineisto oli käsiteltävissä tietokoneen avulla.

Aineistona Biberin tutkimuksessa oli liki miljoonan saneen tekstikorpus, josta pääosa oli koostettu kahdesta laajasta tietokonekorpuksesta. Toinen tekstiaineisto (Lancaster – Oslo – Bergen Corpus) sisältää painettuja tekstejä: mm. kaunokirjallisia tekstejä, eri alojen tieteellisiä tekstejä, lehtitekstejä. Toinen laaja aineisto (London - Lund Corpus) sisältää puhuttua kieltä, yksityisiä ja julkisia keskusteluja, puhelinkeskusteluja, radiopuhetta, spontaaneja ja valmisteltuja puheita. Lisäksi Biber täydensi aineistoa erinäisillä painamattomilla teksteillä kuten yksityiskirjeillä ja tiedeyhteisön sisäisillä kirjeillä. Kaikkiaan tekstejä oli 481, ja Biber kuvaa niiden jakautuvan 23 genren eli tekstilajin kesken. Genrejä ovat hänen mukaansa esimerkiksi tieteelliset tekstit, uskonnolliset tekstit, romanttinen kaunokirjallisuus, lehtireportaasit, radiopuhe, puhelinkeskustelut; nämä puolestaan jakautuvat alagenreihin siten, että tieteellisten tekstien alalajeja ovat mm. lääketieteen, matematiikan, tekniikan,

humanististen alojen tekstit ja lehtitekstien alalajeja mm. urheilua, kulttuuria, taloutta käsittelevät tekstit.

Biber lähti tutkimuksessaan liikkeelle nimenomaan kielen morfologisista, leksikaalisista ja syntaktisista piirteistä, ei kielen funktioista eli merkityksistä tai tilanteisista käyttötavoista. Kuvattavia piirteitä oli kaikkiaan 67. Nämä koskevat erilaisia leksikaalis-kieliopillisia rakenteita, mm. aikamuotoja, ajan ja paikan adverbiaaleja, pronomineja, nominaalistuksia, passiivia, verbityyppejä (statiivinen vs toiminnallinen), alistuksia, modaaliverbejä, rinnastuksia, tyyppien ja esiintymien suhteita (Biber 1988: 71 – 75).

Kun tekstien piirteet oli poimittu ja tunnistettu, Biber selvitti ns. faktorianalyysin avulla, mitkä piirteet esiintyvät yhdessä eli millä tavoin piirteet liittyvät kimpuiksi. Menetelmässä tutkija määrittelee haluamansa piirteet ja antaa koneen löytää tilasto- ja todennäköisyyslaskelmien avulla sen, mitkä piirteet esiintyvät yhdessä ja millä todennäköisyydellä. Kielentutkijan kannalta olennaista on edetä koneen laatimien piirrekimppujen perusteella, joten tekstien kuvaus ei enää ole yhden piirteen varassa – esimerkiksi passiivin esiintymisen tai esiintymättömyyden, puheaktipronominien, tiettyjen aikamuotojen tai nominaalistusten määrän – vaan piirteiden muodostamien kokonaisuuksien. Nämä piirrekokonaisuudet muodostavat jatkumoita – dimensioita – niin, että tekstit ovat jollain kohdalla tätä jatkumoa. Kyse ei siis ole piirteestä, joka tekstillä joko on tai jota ei ole.

Faktorianalyysin jälkeen alkoi tutkimuksen kvalitatiivinen osa, jossa selvitettiin, millaisia dimensioita yhdessä esiintyvät piirteet muodostavat ja millaisia funktioita nämä piirrekimput toteuttavat. Tulosten tulkinnassa Biber nojautui hypoteesiin, että leksikaalis-kieliopilliset piirteet paljastavat systemaattisesti niiden alla olevat funktionaaliset rakenteet eli merkitykset. Kannattaa kuitenkin huomata, että Biber ei sitoudu kielellisen muodon ja merkityksen aukottomaan yhteyteen vaan muotojoukkojen ja merkitysten yhteyteen. Analyysi toi esiin kuusi dimensiota: puhuttavakeskeisyys vs puhujakeskeisyys (*informational versus involved production*), narratiivisuus vs ei-narratiivisuus, eksplisiittisyys vs referenssin tilannesidonnaisuus, argumentoivuus (vaikuttavuus), abstraktius vs ei-abstraktius, tuotoksen reaaliaikaisuus (*on-line informational elaboration*). Tällaiset dimensiot ovat jatkumoita, ja kaikki tekstilajit asettuvat kullekin jatkumolle toisiinsa nähden järjestykseen.

Dimensioiden analyysin jälkeen Biber kuvasi tekstilajit useiden dimensioiden yhteisvaikutuksen perusteella, ei dimensio kerrallaan, eikä varsinkaan yksittäisten kieliopillisten piirteiden perusteella. Menetelmää kuvaa hyvin Biberin käyttämä nimitys: *multi-feature/multi-dimimensional analysis* (MF/MD; suomeksi ehkä monipiirremonidimensioanalyysi).

Kun dimensiot ja niiden kombinaatiot oli selvitetty, tutkimuksessa päästiin varsinaiseen tavoitteeseen: tekstilajien välisten erojen ja tekstilajien sisäisen vaihtelun kuvaukseen. Tulosten tulkinnan tuli perustua nimenomaan dimensioiden yhteisvaikutukseen. Niinpä kahta genreä ei voinut pitää samanlaisena, vaikka ne olivat kahden dimension suhteen samanlaiset, jos ne puolestaan erosivat kahdessa muussa dimensiossa. Esimerkiksi kaksi genreä, keskustelut ja yksityiskirjeet, muistuttavat kaikkien kuuden dimension suhteen toisiaan; spontaanit puheet (monologit) muistuttavat kolmen dimension suhteen keskusteluja: kumpikin genre on melko puhujakeskeinen, ei-abstrakti ja vähän vaikuttava.

Tekstilajien sisäisiä vaihteluja löytyy hyvinkin paljon, mikä johtuu siitä, että joissakin genreissä on runsaasti alalajeja. Lehtitekstin alagenret eroavat useiden dimensioiden suhteen toisistaan; koko lehtigenre on hyvin inkoherentti genre. Radiopuhe puolestaan vaikuttaa hyvin koherentilta genreltä: alagenret eroavat systemaattisesti vain yhden dimension suhteen ja kohtalaisesti erään toisen suhteen, muut neljä dimensiota ovat identtiset.

Kirjansa viimeisen luvun Biber on otsikoinut »Towards a typology of English texts»: käsillä oleva tutkimus antaa mahdollisuuden luokitella tekstit kielellisten piirteiden samanlaisuuden perusteella (s. 206). Typologioinnin tavoitteena on kuvata tekstit kielellisten rakennepiirteiden joukoksi tässä esiteltyjen dimensioiden avulla. Alustavien tietojen mukaan tuo tutkimus monipuolistaa antoisasti sitä tekstitypologiaa, joka on jo vuosikausia ollut vakiintunutta tekstilingvististä yhteistietoa, nimittäin tekstien jaotteleminen narratiiviseen, deskriptiiviseen, ekspositoriseen, argumentoivaan ja instruktiiviseen tyyppiin (Werlich 1979).

Mikä on »piirre» tekstin kuvauksessa?

Jos tavoitteena on kuvata nimenomaan tekstejä, on tällaisessa kvantitatiivisessa tutkimuksessa pohdittava tarkkaan sitä, mitkä piirteet ovat olennaisia juuri tekstien kuvauksessa, ei morfologisten tai syntaktisten piirteiden kuvauksessa. Onko tekstien kuvauksessa olennaista kerätä kaikki mahdolliset poimittavat morfosyntaktiset piirteet vai vain jotkin? Teksteistä löytyy loputtomiin mitattavia piirteitä, joten piirteiden mekaaninen kuvaaminen paljastaa vain osatotuuden; näin arvioivat kvantitatiivista tutkimusta mm. Leech ja Short (1981: 43 – 44). Ts. laskemalla erilaisten piirteiden esiintymiä saadaan tietoa vain näistä piirteistä; se, mikä on piirteiden tehtävä ja tulkinta ja mitä teksti on, jää kuvaamatta. Mikä siis on tekstin piirre? Onko mikä tahansa tunnistettava morfologinen, morfosyntaktinen tai leksikaalinen piirre olennainen tai edes informatiivinen tekstin kuvauksessa?

Toinen keskeinen kysymys on, onko jokaisella morfologisella tai syntaktisella piirteellä yksi tulkinta teksteissä. Suomen kielestä hyvän esimerkin tarjoaa passiivi. Morfologisesti passiivi on vaivatta tunnistettavissa, mutta passiivin lukuisat tehtävät ja käyttötavat eivät ole suinkaan kaikki samanlaisia, kuten seuraavat esimerkit osoittavat:

1) Juodaan kahvia.
2) Mennäänpäs nyt.
3) Suomessa juodaan paljon kahvia.
4) Kahvia säilytetään umpinaisessa pakkauksessa.
5) Tässä artikkelissa käsitellään – –.
6) Juotuaan kahvin mies lähti.

Esimerkissä 1 voi olla kysymys kehotuksesta, jossa puhuja kuuluu mukaan ryhmään, esimerkissä 2 samoin on kehotus, mutta todennäköisesti puhuja ei kuulu mukaan ryhmään. Kolmannessa lauseessa on kyse habitiivisesta deskriptiosta, neljäs voi olla normin tai ohjeen kaltainen, viidennessä passiivimuoto on tekstilajin konventionaalistama (jossain toisessa tekstilajissa olisi sen sijasta tunnusmerkitöntä viitata yksikön ensimmäiseen persoonaan), kuudennessa on lauseenvastikkeessa pelkästään morfologinen passiivi. Jos siis poimimme näkyviin morfologisin kriteerein piirteitä, mitä saamme esiin? Olisiko kuvattavia piirteitä valittaessa otettava huomioon jo passiivin erilaiset tehtävät ja luokiteltava poimittavat muodot esimerkiksi siten, että lauseenalkuinen passiiviverbi on eri asia kuin nominaalilauseketta seuraava passiiviverbi, verbin nominaalimuodon tai modaaliverbin passiivi vielä eri asia? Antaisiko tämä aiheen päätellä, että poimittavaa piirrettä ei voisi rajata morfologisesti vaan vain funktion mukaan? Ts. Biberin lähtökohta poimia avoimin mielin morfosyntaktisia piirteitä sellaisinaan, koota sitten piirteet konevoimin sokeiksi faktoreiksi ja vasta tämän jälkeen tulkita funktiot voi johtaa sattumanvaraisiin dimensioihin ja siihen, että monifunktioiset morfeemit ovat merkityksettömiä dimensioiden muodostumisessa, koska ne näyttäisivät kombinoituvan hyvin monenlaisten piirteiden kanssa eli tilastolaskujen mukaan esiintyvän sattumanvaraisesti.

Kolmannen pulman tekstin piirteiden poimintaan aiheuttaa se, että tekstin kuvauksen kannalta on suuri joukko olennaisia piirteitä, joihin ei mekaanisesti poimimalla päästä käsiksi. Ulkopuolelle jää mm. tekstin retoriikka eli se, miten kirjoittaja jäsentää sanottavansa lineaariseksi teks-

tiksi, käsitteiden ja tarkoitemaailman suhteet, kirjoittajan ja lukijan vuorovaikutus sekä toisaalta kirjoittajan suhtautuminen topiikkiin. Olennaisia ovat myös erilaiset »tekstin näkökulmaa» muokkaavat seikat, joista vain osa voidaan tavoittaa mitattavilla ja tunnistettavilla leksikaalis-kieliopillisilla piirteillä (tästä ks. Fowler 1986). Näitä kuvattaessa on tarkasteltava nimeämistä, analysoitava käytettyjä käsitteitä ja niiden semanttisia yhteyksiä, tekstin virkkeiden ja lauseiden suhteita toisiinsa ym. Kriittisten lingvistien mielestä – joita mm. Fowler edustaa – kaikki tekstin piirteet ovat kirjoittajan valinnan tulosta, ja ne ovat siten merkityksellisiä kulloisessakin tekstissä. Piirteet on mahdollista tulkita vain tekstikokonaisuutta tarkasti lukemalla ja analysoimalla. (Fowler 1986: 53 – 62.)

Neljänneksi vaatii tarkkaa piirteiden analyysiä havaita, ovatko löydökset juuri niitä, mitä niiden uskotaan olevan. Entäpä jos piirteiden korrelaatio johtuukin jostakin toistaiseksi tavoittamattomasta piirteestä? Esimerkkinä tällaisesta Myhill esittää tutkimuksessaan (1992: 258 – 259) korrelaation, jonka mukaan espanjankielisen lauseen sanajärjestys pyrki olemaan tyyppiä SV, jos pronominin ja korrelaatin etäisyys oli pienin mahdollinen. Tarkempi tarkastelu osoitti kuitenkin, että sanajärjestys oli kytköksissä pronominin subjektiasemaisuuteen eikä viittaukseen läheisyyteen.

Tekstin kuvaus tällaisella »bottom-up»-menetelmällä redusoi tekstin leksikaalisten ja morfosyntaktisten piirteiden kombinaatioksi. Tämän kannanoton Biber tekee tietoisesti: ensisijaisia ovat hänen mukaansa kielelliset piirteet (so. muodot, rakenteet), funktionaaliset ovat toissijaisia (s. 12 – 13). Näin on mahdollista kuvata esimerkiksi typologisia tendenssejä ja erotella yleisiä ja harvinaisia rakenteita sekä saada tietoa rakenteiden esiintymisympäristöistä. Myhillin mukaan määrittelemällä leksikaalis-kieliopillisten piirteiden esiintymisympäristöt saadaan kuvatuksi näiden merkitykset (1992: 257). Tällöin tuloksena ei kuitenkaan ole varsinainen tekstianalyysi, vaan tutkimus koskee tekstilauseiden leksikaaliskieliopillisia piirteitä. Tällaiset tutkimukset ovat tuottaneet runsaasti soveltamiskelpoista tietoa mm. kielenopetukseen. Esimerkiksi opetettaessa tieteellisten englanninkielisten tekstien laatimista ulkomaalaisille on opetuksen sisältöjä voitu painottaa nimenomaan sen mukaan, mitkä rakenteet ovat todennäköisiä kyseisissä tekstilajeissa (Swales 1990: 2).

Mikä on genre?

Biber väittää tarkoittavansa genrellä tekstijoukkoa, joka on määräytynyt kielenulkoisten kriteerien perusteella puhujan tai kirjoittajan tarkoituksen mukaan (s. 68). Kuitenkin hänen aineistonsa on luokiteltu toisin: genret määräytyvät topiikin (esim. uskonnolliset tekstit), välineen (radiopuhe, puhelinkeskustelu) tai yhteisön (esim. tieteelliset tekstit) mukaan. Alagenret puolestaan määräytyvät joko topiikin mukaan tai sosiaalisen toiminnan mukaan. Esimerkiksi lehtitekstit jakautuvat topiikiltaan politiikkaa, kulttuuria, taloutta käsitteleviin, tieteelliset tekstit taas matematiikkaa, lääketiedettä, yhteiskuntatieteitä käsitteleviin. Sosiaalisen toiminnan mukaan jaoteltuja spontaanin puheen alajaleja ovat esimerkiksi oikeudenistunnossa pidetyt puheet ja päivällispöytäkeskustelut, valmisteltujen puheiden alalajeja ovat mm. saarnat, yliopiston luennot ja yleisöesitelmät. Kun tutkimus tuo tietoa tällä tavoin koostetuista genreistä, mistä se silloin tuo tietoa? Mikä siis on genre? Ja miten tulisi suhtautua tutkimuksen tuomiin tietoihin genrejen välisistä eroista ja genrejen sisäisestä variaatiosta?

Jos yhteen genreen kuuluu akateemisia lääketieteellisiä tekstejä, esimerkiksi tutkimusartikkeleita ja -raportteja, kyseessä on hyvin koherentti genren rajaus. Jos taas toiseen kuuluu lehden urheilutekstejä, esimerkiksi uutinen urheilujohtajan nimityksestä, huippu-urheilijan haastattelu, reportaasi dopingista, pesäpallon sarjapelien tuloksia,

kyseessä on kovin kirjavasti rajattu genre. Eikö tulos ole lopuksi jollain lailla kehämäinen, kun tilastollisen tutkimuksen avulla saadaan selville, että urheilugenren sisällä on suurempi vaihtelu ja hajonta kuin lääketieteen genren?

Millä tavoin tekstit siis tulisi niputtaa, jotta nuo niput olisivat tekstien kannalta olennaisia? Arkipäivän leimat tuskin riittävät, malliin »lehtireportaasi», »yleisönosastonkirjoitus», »tieteellinen artikkeli». Tekstin muotoa ja rakentumista ei ilmiselvästi määrää niinkään topiikki kuin puhujan/kirjoittajan tarkoitus ja yhteisö, jossa teksti kirjoitetaan (tästä ks. erityisesti Swales 1990) sekä puhujan/kirjoittajan ja kuulijan/lukijan välisen kontaktin laatu (ei kontaktia – yksisuuntainen kontakti – kaksisuuntainen kontakti; ks. Martin 1992). Yhteisö ja tarkoitus määräävät kirjoittajan ja lukijan roolin, sen miten asiaa käsitellään ja sen millainen tekstin kielellinen asu lopulta on. Esimerkiksi jos tavoitteena on tutkimustulosten selvittäminen ja päätelmien perustelu ja diskurssiyhteisönä joukko tutkijoita, ovat tekstin useat kielelliset valinnat määräytyneet jo näillä perusteilla, olipa topiikki mikä tahansa. Niinpä jos kieltä tarkastellaan toimintana, aineistokin tulisi niputtaa toiminnoittain, ei topiikeittain tai välineittäin. Toistaiseksi tällaisen kvantitatiivisen tutkimuksen pohjaksi näyttäisi olevan hyvin vähän genreanalyyttisiä tekstin kuvauksia eli systemaattisia kuvauksia siitä, miten tekstit eriytyvät toiminnoittain.

Uudehkossa artikkelissaan (1992) Biber on irtautunut genrejen käsittelystä; hän ei puhu genreistä eikä tarkastele tekstejä yhtenäisenä ryhmänä tällä perusteella, ts. hän ei enää tarkastele tekstejä toiminnoittain vaan rakenteittain eriytyneinä. Terminologisesti tämä näkyy siinä, että genren sijasta hän puhuu rekistereistä. Kyse ei kuitenkaan ole pelkästään terminologiasta vaan tarkastelunäkökulma on toinen: tekstiä tarkastellaan rakenteellisena yksikkönä, kielellisten valintojen joukkona, rekisterinä, eikä sosiaalisen toiminnan määräämänä tilanteisesti varioivana kielenkäyttönä (genren ja rekisterin erosta ks. Swales 1990: 40 – 41; Martin 1985: 240). Rekisterien typologinen analyysi antaa kiinnostavaa tietoa sekä kieliopillisista rakenteista että kielenkäytöstä, mutta tällainen tutkimus ei pysty eikä edes pyri tavoittamaan sitä, mikä on teksti eli mitä on kielenkäyttö ja miksi tekstit eroavat toisistaan, mikä on tekstin merkitys, miten tekstin tulkinta syntyy, miten yksittäisten ilmausten merkitys määräytyy kokonaisuuden mukaan, millaisessa dialogissa teksti on mukana ja miten tuo dialogi vaikuttaa tekstin rakenteeseen ja tulkintaan.

PIRJO KARVONEN

LÄHTEET

BIBER, DOUGLAS 1988: Variation across speech and writing. Cambridge University Press. Cambridge.
1992: On the Complexity of Discourse Complexity: A multidimensional Analysis. Discourse Processes 15: 133 – 164.

FOWLER, ROGER 1986: Linguistic Criticism. Oxford University Press. Oxford.

LEECH, GEOFFREY N. – SHORT, MICHAEL H. 1981: Style in fiction. Longman. London.

MARTIN, JAMES 1985: Process and text: two aspects of human semiosis. – Benson, James D. – Greaves, William S. (toim.): Systemic perspectives on discourse. Vol. I. Norwood, NJ. Ablex.
1992: English Text. System and Structure. John Benjamins Publishing. Philadelphia.

MYHILL, JOHN 1992: Typological Discourse analysis. Blackwell. Oxford.

SWALES, JOHN 1990: Genre Analysis. Cambridge University Press. Cambridge.

WERLICH, EGON 1979: Typologie der Texte. Heidelberg.