Englannin kielen tutkimusta tietokoneiden aikaan

Matti Rissanen

Tietokonekorpukset ovat mullistaneet etenkin kielen varhaisten vaiheiden tutkimuksen. Korpusten tukema kielen vaihteluun perustuva muutosteoria on myös yhdistänyt nykykielen ja kielihistorian tutkimuksen. Ennen kaikkea aikaisemmin jopa vuosia vienyt esimerkkiaineiston kerääminen ja kirjaaminen voidaan korpusten avulla suorittaa murto-osassa tästä ajasta. On kuitenkin muistettava, että kielentutkijalle korpukset tarjoavat vain tehokkaan menetelmän tutkimusaineiston tallennukseen sekä kielen lukemattomia ilmaisuja valaisevien esimerkkien poimimiseen sekä laadulliseen ja määrälliseen luokitteluun. Aineiston analysointi, tulosten merkityksen havainnointi ja osoittaminen, yleistysten ja teorian kehittely on edelleen tutkijan aivotyön varassa.

Viimeisen puolen vuosisadan aikana on kielentutkimuksessa tapahtunut kaksi merkittävää mullistusta. Ensimmäinen niistä oli amerikkalaisen, poliittisena kannanottajanakin nimeä saavuttaneen Noam Chomskyn 1950-luvulla kehittelemä transformatiivis-generatiivinen kielioppi, joka siirsi kielen selityksen ja analyysin vahvasti abstraktille tasolle. Tämä teoria oli tärkeä kielen systemaattisen kuvauksen kannalta, mutta se ei kiinnittänyt riittävästi huomiota kielen perusluonteeseen kommunikaation välineenä, jonka tarkoitus on välittää viestejä, tietoa ja tunteita ja pitää yllä yhteyttä ihmisten kesken. Kielen perusolemusta ei voi ymmärtää ymmärtämättä ihmistä, ja tätä ei transformaatiokielioppi korostanut kylliksi.

Jo seuraavalla vuosikymmenellä alkoi kehittyä toinen voimakas kielen selitysmalli ja teoria vastapainoksi kielen kommunikaatioluonteen unohtaneelle suuntaukselle. Malli perustuu kielen vaihtelevuuteen, variaatioon, ja se tarkastelee kieltä monimuotoisena kokonaisuutena, ihmisten keskinäisen kommunikaation välineenä. Ilmaisuvaihtoehtojen usein tiedostamatonta valintaa säätelevät mitä moninaisimmat sosiaaliset, alueelliset sekä viestin tarkoitukseen ja viestintätilanteeseen liittyvät seikat. Vaihtoehdot näkyvät tietenkin luontevimmin sanastossa, mutta ne näyttelevät tärkeää osaa myös kieliopin rakenteissa ja ääntämisessä, ja jopa yleisten tekstistrategioiden valinnassa. Tämän vaihtelun johdosta kieli muuttuu jatkuvasti – eilen, tänään ja huomenna. Jotkin ilmaisuvariantit katoavat ja uusia syntyy koko ajan, joskin muutokset ovat yleensä helpoimmin havaittavissa joidenkin varianttien yleistymisenä ja toisten käytön vähenemisenä. Muutosta ei siis pidä mieltää pölyiseksi ja epäkiinnostavaksi menneisyyden ilmiöksi, vaan jännittäväksi jatkuvaksi tapahtumaksi, jonka keskellä elämme, ja josta meidän on syytä olla koko ajan tietoisia.

Vaihtelun aiheuttajia

Ajatus kielen vaihtelusta ja vaihteluun perustuvasta muutoksesta ei tietenkään ole uusi; jo vuosisatojen ajan on huomattu, että eri ihmiset puhuvat samaa kieltä eri tavoin, ja että joidenkin ihmisten ja ihmisryhmien käyttämä kieli tuottaa muutoksia enemmän kuin toisten. Teoreettiseksi rakennelmaksi tämä tutkimuslähtökohta hahmottui kuitenkin vasta 1960- ja 1970-lukujen taitteessa. Teorioiden kehittäjistä huomattavimpia olivat amerikkalaiset Uriel Weinreich, William Labov ja Marvin I. Herzog sekä nykyisin Australiassa vaikuttava M. A. K. Halliday.

Hallidayn mukaan kielen perustana on merkityspotentiaali, lukematon määrä merkityksiä. Kutakin merkitystä voidaan ilmaista suurella joukolla kielellisiä ilmaisuvaihtoehtoja, variantteja. Suurin piirtein samaa merkitsevät ilmaisuvaihtoehdot voidaan ryhmittää varianttikentik-

si. Varianttikenttien sisällä ilmaisuvaihtoehdon valintaan vaikuttavat monet eri tekijät, joiden kartoittaminen ja selittäminen on kielentutkijan tärkeimpiä tehtäviä. Täysin kattavan ja aukottoman kuvauksen laatiminen näistä tekijöistä olisi luonnollisesti ylivoimainen tehtävä, mutta yleisiä suuntaviivoja ja tekijäkimppuja on melko helppo määritellä ja rajata. Esimerkiksi seuraavat tarkastelun kohteet ovat keskeisiä variaatiotutkimukselle:

Alueellinen vaihtelu
– Paikallismurteet- Paikallismurteet
– Brittienglanti, amerikanenglanti jne.- Brittienglanti, amerikanenglanti jne.

Sosiolingvistinen vaihtelu
– Puhujan ja kuulijan välinen suhde- Puhujan ja kuulijan välinen suhde
– Puhujan sosiaalinen asema, koulutus jne.- Puhujan sosiaalinen asema, koulutus jne.

Tekstilajin aiheuttama vaihtelu
– Tieteen kieli, sanomalehtikieli, yksityiskirjeet, reseptit, käyttöohjeet jne.- Tieteen kieli, sanomalehtikieli, yksityiskirjeet, reseptit, käyttöohjeet jne.

Puhuttu kieli / kirjoitettu kieli

Paikallismurteet ovat kaikkien tuntema selitys sille, miksi samankin kielen puhujat voivat käytellä kieltä varsin eri tavoin. Ajatellaanpa vaikka kotimurrettaan puhuvaa savolaista, hämäläistä tai varsinaissuomalaista. Englannin kielen alueelliseen variaatioon kuuluvat tietenkin myös eri puolilla maailmaa puhutut englannit: Ison Britannian ja Yhdysvaltain lisäksi vaikkapa Australiassa, Kanadassa tai Etelä-Afrikassa käytetyt kielivarieteetit (ks. Nevalaisen ja Pahdan artikkelit tässä lehdessä).

Sosiolingvistinen vaihtelu on paljon monitahoisempi asia (ks. Raumolin-Brunbergin artikkeli tässä lehdessä). Tämän vaihtelun piiriin kuuluu muun muassa puhujan ja kuulijan välinen suhde, joka määräytyy esimerkiksi heidän sosiaalisesta asemastaan tai siitä, kuinka hyvin he tuntevat toisensa; todennäköisesti puhumme tai kirjoitamme esimerkiksi tasavallan presidentille aivan eri tapaan kuin kollegoille tai perheenjäsenille.

Myös tekstilajin ja viestintätilanteen aiheuttama variaatio on voimakas ja monitahoinen vaihtelutekijä. Tieteen kielellä on oma erityinen ilmaisutapansa (ks. Taavitsaisen artikkeli tässä lehdessä). Sanomalehdillä on oma tyylinsä, samoin resepteillä ja käyttöohjeilla, puhumattakaan yksityiskirjeistä ja erilaisista kaunokirjallisista teksteistä. On myös muistettava, että kirjoitettu kieli eroaa monessa suhteessa puhutusta kielestä.

Edellä mainittujen seikkojen lisäksi ei tietenkään pidä unohtaa myöskään sitä, että meillä jokaisella on oma yksilöllinen tapamme ilmaista itseämme. Toiset puhuvat ja kirjoittavat selkeän asiallisesti ja niukkasanaisesti välttäen korostuksia; toisten kieli taas on värikästä, painokasta, kuvailmaisuja viljelevää. Lisäksi ihmisten kyky ja taito hallita useita eri ilmaisun tasoja, rekistereitä, vaihtelee suuresti esimerkiksi perhetaustasta ja koulutuksesta riippuen.

Vaihtelusta muutokseen

Ehkä jo edellä esitetystä suppeasta ja väistämättä ylimalkaisesta hahmottelusta voi päätellä, että vaihteluun perustuva kielentutkimus on monitahoista ja tieteidenvälisyyttä edellyttävää työtä. Mutta vielä haastavammaksi se muuttuu, kun siihen liitetään kielen muutoksen tutkimus. Variaatioteorian keskeisiä teemoja on, että kieli muuttuu suureksi osaksi vaihtelun perusteella. Aikaisemmin muutoksen kuvaustapa oli suunnilleen tällainen:

A > B

Jokin kielen ilmaisutapa siis muuttuu toiseksi, ja sillä hyvä. Hieman karrikoiden voidaan lisätä, että jos tällaisen mallin esittäjältä erehtyi kysymään, mistä sitten johtui, että A muuttui B:ksi, saattoi tämä vastata: ”Koska on sääntö, että A:sta tulee B.”

Tämä muutoksen kuvaus on tietenkin selkeä ja saattaa palvella riittävästi jonkinasteisia teoreettisia ajatusrakennelmia kielen muutoksesta. Sen vikana on kuitenkin, että se antaa puutteellisen kuvan kielen muutoksen todellisuudesta ja siihen vaikuttavista tekijöistä. Jos halutaan havainnollistaa muuttuvaa ilmaisua isoin kirjaimin, lähempänä kielen todellisuutta on seuraavanlainen kuvaustapa:

A	A
A	A
A	B
A	B
A	B
A	B
B	C
B	C

Toisin sanoen kielen muutos ei suinkaan aina tai edes useimmiten merkitse sitä, että jokin ilmaisuvaihtoehto katoaa ja toinen tulee tilalle, vaan paljon useammin ja tyypillisemmin muutoksia varianttikentän ilmaisuvaihtoehtojen yleisyydessä ja keskinäisissä suhteissa. A:sta ei siis tule B, vaan ilmaisutapa A, joka on ollut yleisin varhaisempana aikana, menettää asemaansa suunnilleen samaa merkitsevälle ilmaisulle B, ja jostakin ilmestyy uusi vaihtoehto C. Väistyvät variantit leimautuvat usein tyylillisesti – vanhanaikaisiksi, juhlallisiksi tai vain erikoistilanteisiin sopiviksi. Vähitellen ne saattavat tietenkin kadota kokonaan mutta usein vasta pitkän prosessin seurauksena.

Kielen muutoksen tutkija pyrkii selvittämään, mitkä tekijät vaikuttavat varianttikentän yllä kuvattuun muutokseen, aivan samoin kuin hän vaihtelua tutkiessaan kysyy, mitkä tekijät vaikuttavat tietyn ilmaisuvaihtoehdon valintaan. Tällöin tutkimuksen monitieteisyys korostuu. Muuttuva yhteiskunta, esimerkiksi agraarisesta teollistuneeksi, luokkarajojen syntyminen ja katoaminen, tieteen ja kaunokirjallisuuden kehityslinjat, eri kielten ja murteiden puhujien liikkuvuus ja uudet kontaktit vaikuttavat monien muiden tekijöiden ohella sanojen ja rakenteiden syntymiseen ja vanhojen väistymiseen.

Kielen ilmaisut muuttuvat jatkuvasti myös niin kutsuttujen kielensisäisten muutosprosessien johdosta. Meidän puheemme on, jälleen yksinkertaistaen, jatkuvaa tasapainoilua pienimmän ponnistelun ja ymmärretyksi tulemisen tavoitteiden välillä Etenkin puhuessamme yritämme lyhentää ja yksinkertaistaa ilmaisuamme erityisesti ääntämisen mutta myös kieliopin tasolla. Sanomme mieluummin ”tuu” kun ”tule” ja ”mennään” kuin ”menkäämme” tai ”me menemme”. Samalla kertaa meidän on kuitenkin huolehdittava siitä, ettei puheemme muutu täysin käsittämättömäksi muminaksi, etteivät tärkeät merkityserot katoa, ja että myös ilmaisun persoonallisuus ja ilmeikkyys ainakin jossakin määrin säilyvät. Toistensa kanssa kilpailevia ilmaisuvaihtoehtoja syntyy myös tällä tavoin, ja esimerkiksi puhetilanteen muodollisuuden taso ja puhujien sosiaaliset erot vaikuttavat merkittävästi kielensisäisten muutosprosessien aiheuttamaan vaihteluun.

Tietokone kielentutkijan apuna

Kielen vaihtelun tutkimus vaatii luonnollisesti suurten tekstimäärien hallintaa. Esimerkkejä ilmaisuvaihtoehdoista on löydettävä runsaasti, jotta voitaisiin tehdä edes kohtuullisen luotettavia päätelmiä niiden yleisyydestä, keskinäisistä suhteista ja ennen kaikkea niiden valintaan vaikuttavista tekijöistä. Kielen muutosta tutkittaessa tekstimateriaalin runsauden vaatimus yhä vain kasvaa. Vaihtelua on kartoitettava toisiaan seuraavina aikakausina muutoksen seuraamiseksi ja havaitsemiseksi. Lisäksi kirjoitetut tekstit ovat ainoa tapa saada ensikäden tietoa menneiden vuosisatojen tai jopa vuosituhansien kielestä. Englanninkielisiä tekstejä on kirjoitettu lähes puolentoista vuosituhannen ajan, klassillisten kielten kohdalla aikajänne on tietenkin vielä paljon pitempi.

Tietokoneteknologia on jo muutaman vuosikymmenen ajan tarjonnut korvaamatonta tukea kielentutkijalle tekstien tallennuksessa sekä esimerkkien etsinnässä ja analysoinnissa. On luotu niin kutsuttuja tietokonekorpuksia, laajoja elektronisesti tallennettuja tekstikokoelmia, joiden tekstit on huolellisesti valikoitu ja järjestetty siten, että ne antavat joko hyvän yleiskuvan kielestä tai keskitetyn kartoituksen jostakin kielen alueesta, kuten yksityiskirjeistä, tieteen kielestä tai murteista.

Korpuksia on koottu sekä nykyenglannista että englannin kielen varhaisemmista vaiheista. Suurimmat nykyenglannin korpukset käsittävät satoja miljoonia sanoja; pienimmät, kohdennetut korpukset voivat olla muutaman sadan tuhannen sanan suuruisia. Nykyenglannin korpukset kattavat myös alueellisia varieteetteja (britti- ja amerikanenglantia, Australian englantia, Itä-Afrikan englantia, Intian englantia, Uuden Seelannin englantia ja niin edelleen).

Nykyenglannin korpuksia

British National Corpus (n. 100 milj. sanaa)
Bank of English (n. 320 milj. sanaa)
Brown Corpus (amer.engl. 1960-luv. 1 milj. sanaa)
Frown Corpus (amer. engl. 1990-luv. 1 milj. sanaa)
LOB Corpus (brittiengl. 1960-luv. 1 milj. sanaa)
FLOB Corpus (brittiengl. 1990-luv. 1 milj. sanaa)
International Corpus of English (ICE) (1 milj. sanaa/korpus)
- East Africa
- Great Britain
- India
- New Zealand
- Philippines
- Singapore

Myös internetiä on alettu käyttää systemaattisesti englannin korpuspohjaisessa tutkimuksessa.

Englannin kielen korpuksia

Helsingin yliopiston englannin kielen laitoksen yhteydessä toimiva Englannin kielen vaihtelun ja muutoksen tutkimusyksikkö sai alkunsa 1980-luvulla tutkimusprojektina, jonka tuloksena valmistui 1990-luvun alussa Helsingin korpus, *The Helsinki Corpus of English Texts.* Sen laadintaan osallistuivat muun muassa kaikki tämän numeron anglistiartikkeleiden kirjoittajat sekä useat muut englannin kielen laitoksen nuoret ja varttuneet tutkijat. Projektiryhmän sihteeri oli nykyisin Uppsalan yliopiston professorina toimiva Merja Kytö. Helsingin korpus oli ensimmäinen yritys kartoittaa englannin kielen menneisyys monipuolisen tekstivalikoiman muodossa pitkällä aikakaarella, 700-luvulta 1700-luvun alkuun, ja se on yhä käytössä sadoissa yliopistoissa ja tutkimuskeskuksissa eri puolilla maailmaa.

Peruskorpuksen valmistuttua alkoi 1990-luvulla syntyä ”toisen sukupolven korpuksia”, kuten Anneli Meurman-Solinin kokoama varhaisen skotin korpus, Terttu Nevalaisen ja Helena Raumolin-Brunbergin ryhmän kokoama varhaisten yksityiskirjeiden korpus (ks. Raumolin-Brunbergin artikkeli tässä numerossa) ja Irma Taavitsaisen ja Päivi Pahdan ryhmän lääketieteen englannin korpus (ks. Taavitsaisen artikkeli). Loppusuoralla on myös brittienglannin nykymurteiden korpusprojekti, jonka edesmennyt professori Tauno F. Mustanoja aloitti jo 1970-luvulla, ja jota on viime vuodet johtanut dosentti Kirsti Peitsara.

Tämän vuosikymmenen puolella on aloitettu myös ”kolmannen sukupolven” tarkoin kohdennettujen historiallisten korpusten laadinta. Tutkimusyksikkömme nuorten jäsenten valmisteilla olevista korpuksista voidaan mainita varhaisten kasvitieteellisten tekstien korpus (*Martti Mäkinen*), 1700-luvun kirjeitä sisältävä Bluestocking Corpus (*Anni Sairio*), seuraavalla vuosisadalla eläneiden köyhien ihmisten kirjeiden korpus (*Mikko Laitinen*) sekä 1500- ja 1600-luvuilla kirjoitettujen noitapamflettien korpus (*Carla Suhr*) ja tupakkapamflettien korpus (*Maura Ratia*). Kansainvälisiä yhteistyöprojekteja ovat 1600-luvun lopun Uutta Englantia kuohuttaneiden Salemin noitavaino -oikeudenkäyntien dokumenttien korpus sekä 1600-luvulta meidän päiviimme ulottuva brittiläisten ja amerikkalaisten tekstien yleiskorpus, joka jatkaa Helsingin korpuksen aikajännettä.

Luonnollisesti myös muualla maailmassa laaditaan suuria ja kunnianhimoisia historiallisia korpuksia. Englannin kielen varhaisimmista vaiheista on sanakirjaprojektien yhteydessä kerätty valtavia tekstitiedostoja, joskin ne ovat juuri koostaan johtuen vähemmän organisoituja ja strukturoituja kuin Helsingin korpus. Edelleen esimerkiksi kaikki Shakespearen näytelmät on koottu korpukseksi, ja kaupallisten tietopankkien puolelta löytyy erittäin suuri määrä kaunokirjallisia tekstejä, etunenässä edustava mutta kallis LION, eli Literature Online.

Ensimmäiset nykyenglannin ja englannin kielen historian korpukset eivät vielä käyttäneet kovinkaan paljon hyväkseen kaikkia tietokonetekniikan mukanaan tuomia mahdollisuuksia. Helsingin korpus oli uranuurtava antaessaan jokaisen tekstiotteen alussa kohtuullisen tarkan tiedon paitsi tekstin iästä, murteesta ja tekstilajista myös mahdollisuuksien mukaan kirjoittajan iästä, sukupuolesta ja sosiaalisesta asemasta. Tässä suhteessa kehitys on edennyt huikeasti muutamassa kymmenessä vuodessa, ja tekstianalyysiin perustuva tutkimus on ratkaisevasti helpompaa ja nopeampaa kuin ennen. Monien korpusten tekstit on koodattu kieliopillisesti, joten tekstiä voi tutkia paitsi sanojen ja niiden osien tai yhdistelmien perusteella myös etsimällä sanaluokkia tai lauseenjäseniä ja niiden yhdistelmiä. Tulevaisuuden korpusvisioistamme voidaan vielä mainita todelliset multimediakorpukset, joissa tekstiin liittyy myös alkuperäiskäsikirjoitus, kuvia, monenlaista taustatietoa ja tarpeen mukaan myös ääntä. Sanakirja- ja sanastotiedon lisääminen on myös mahdollista.

Kielitieteilijät ovat varsin yksimielisiä siitä, että tietokonekorpukset ovat mullistaneet täysin etenkin kielen varhaisten vaiheiden tutkimuksen. Korpusten tukema kielen vaihteluun perustuva muutosteoria on myös yhdistänyt nykykielen ja kielihistorian tutkimuksen. Ennen kaikkea aikaisemmin jopa vuosia vienyt esimerkkiaineiston kerääminen ja kirjaaminen voidaan korpusten avulla suorittaa murto-osassa tästä ajasta. Mutta samanaikaisesti on muistettava, että kielentutkijalle korpukset tarjoavat vain tehokkaan menetelmän tutkimusaineiston tallennukseen sekä kielen lukemattomia ilmaisuja valaisevien esimerkkien poimimiseen sekä laadulliseen ja määrälliseen luokitteluun. Aineiston analysointi, tulosten merkityksen havainnointi ja osoittaminen, yleistysten ja teorian

kehittely on edelleen tutkijan aivotyön varassa. ”Tutkimus alkaa siitä, mihin laskeminen loppuu”, onkin yksi tutkimusyksikkömme mottoja.

KIRJALLISUUTTA

Halliday, M. A. K. (1973): *Explorations in the Functions of Language,* Edward Arnold, London.

Rissanen, Matti (2000): ’The World of English Historical Corpora: from Cædmon to Computer Age’, *Journal of English Linguistics* 28: 7-20.

Rissanen, Matti, Merja Kytö and Minna Palander-Collin (toim.)(1993), *Early English in the Computer Age: Explorations through the Helsinki Corpus*. Mouton de Gruyter, Berlin & New York.

Weinreich, Uriel, William Labov and Marvin Y. Herzog (1968): ’Empirical Foundations for a Theory of Language Change’, teoksessa W. P. Lehmann ja Yakov Malkiel (toim.) (1968): *Directions for Historical Linguistics: a Symposium,* University of Texas Press, Austin, Texas.

Matti Rissanen on englantilaisen filologian emeritusprofessori ja englannin kielen vaihtelun ja muutoksen tutkimusyksikön tutkija Helsingin yliopistossa.