Luku 9 Korpusaineistot

Veronika Laippala
https://orcid.org/0000-0002-7635-429X

Minna Palander-Collin
https://orcid.org/0000-0001-8428-4423

Mikä?

Korpuksella tarkoitetaan kielentutkimuksessa laajaa sähköisessä muodossa olevaa tietokoneluettavaa tekstikokoelmaa, joka on strukturoitu ja edustava tutkimuksen tarpeita palvelevalla tavalla. Korpus voi sisältää erilaisia hyödyllisiä lisätietoja, kuten sanaluokkakoodeja tai lauseenjäsennyksen. Kukin korpuksen tekstikatkelma alkaa yleensä metatiedoilla, joista selviää esimerkiksi tekstin kirjoittaja tai puhujat, tekstilaji ja tekstin tuottamisen ajankohta. Näiden merkintöjen avulla korpuksesta voidaan tehokkaasti tehdä erilaisia automatisoituja hakuja.

Katso myös:

Luku 7 Laadullinen aineistopohjainen kielentutkimus

Luku 10 Määrällinen korpuslingvistiikka

Luku 11 Historiallinen korpuslingvistiikka

1. Johdanto

Korpukset ovat nykyaikaisessa kielentutkimuksessa keskeinen aineistotyyppi, sillä ne tarjoavat tutkijalle empiiristä aineistoa kielenkäytöstä. Niiden avulla voidaan vastata monenlaisiin tutkimuskysymyksiin, testata kielenkäyttöä koskevia hypoteeseja ja tarkistaa esimerkiksi, kuinka tiettyä sanaa käytetään. Tyypillisesti korpuksista saadaan tavalla tai toisella kielestä määrällistä tietoa, jollaista kielen käyttäjien intuitio ei tavoita (ks. Määrällinen korpuslingvistiikka tässä kirjassa). Korpukset soveltuvat kuitenkin myös laadullisten tutkimusten aineistoksi (ks. Laadullinen aineistopohjainen kielentutkimus tk.).

Korpukset ovat tutkijan kannalta nopea keino saada tutkimuksessa tarvittava aineisto, sillä monista kielistä on jo olemassa eri tarkoituksiin koottuja valmiita korpuksia. Oman korpuksen kokoaminenkin voi tuntua helpolta ja houkuttelevalta vaihtoehdolta erityisesti, jos aineisto on jo valmiiksi sähköisessä muodossa esimerkiksi netissä. Korpuksen kokoaminen on kuitenkin aina muutakin kuin vain tekninen ”imurointisuoritus”, ja varsinkin aloittelevan korpustutkijan kannattaa ensin selvittää, millaisia korpuksia on valmiiksi saatavilla. Keskeistä on miettiä, mihin kysymyksiin korpuksella pyritään vastaamaan ja mitä edellytyksiä tämä puolestaan asettaa aineistolle.

Tässä artikkelissa valotamme ensin luvussa 2 lyhyesti korpusten historiaa ja korpustutkimuksen taustalla olevaa kielikäsitystä ja muita tekijöitä. Luvussa 3 esittelemme valmiina saatavilla olevia aineistoja ja luvussa 4 taas tarkastelemme korpusten koostamisessa huomioon otettavia seikkoja. Luvussa 5 esittelemme Kielipankin Korp-käyttöliittymän avulla, miten korpustyökalut toimivat. Lopuksi pohdimme vielä korpustutkimuksen viimeaikaista kehitystä, ns. big datan synnyttämiä uusia metodologisia kysymyksiä ja korpustutkimuksen rajoitteita ja sudenkuoppia.

2. Korpustutkimuksen taustaa

Vaikka periaatteessa mikä tahansa systemaattisesti koottu tekstikokoelma voi muodostaa korpuksen, korpuksella tarkoitetaan yleensä nimenomaan tietokoneluettavaa tekstiaineistoa. Korpuslingvistiikan synty nivoutuukin yhteen tietotekniikan kehityksen kanssa. Tärkeä merkitys on ollut myös kielitieteen kehityksellä, sillä viimeistään 1980-luvulta alkaen on tukeuduttu yhä enenevässä määrin aineistopohjaisiin, empiirisiin menetelmiin muun muassa generatiivisen kielentutkimuksen lähes yksinomaisesti käyttämän introspektion eli kielitajun lisäksi (ks. Kielitaju kielentutkijan työkaluna tk.).

The Brown Corpus of Contemporary American English,[1] joka julkaistiin vuonna 1964 Brownin yliopistossa Yhdysvalloissa, oli ensimmäisiä tietokoneluettavia korpuksia. Se sisältää miljoona sanaa vuonna 1961 kirjoitettua amerikanenglantia ja koostuu 500:sta eri tekstikategorioita edustavasta 2 000 sanan otoksesta. Nykymittapuulla Brown-korpus on varsin pieni, mutta sitä käytetään tutkimuksessa edelleen.

Hyödylliseksi Brown-korpuksen tekee se, että muita korpuksia on koottu vastaavalla sapluunalla, kuten vuoden 1961 brittienglantia edustava Lancaster-Oslo-Bergen (LOB) Corpus[2]. Lisäksi alkuperäisten Brown- ja LOB-korpusten verrokit, 1990-luvun alun englantia edustavat Frown[3] ja F-LOB[4] sekä 1930-luvun englantia edustavat B-LOB-1931[5] ja B-Brown[6], mahdollistavat sekä ajallisen että maantieteellisen vertailun. Tällainen aineistojen yhteismitallisuuden ja vertailtavuuden lisääminen voisi yleisemminkin olla paikallaan tutkimuksen läpinäkyvyyden ja toistettavuuden sekä tiedon kumuloitumisen kannalta.

Myös Suomessa korpustutkimus lähti vauhtiin aikaisessa vaiheessa. Oulun yliopistossa laadittiin jo 1960-luvulla suomen kielen Oulun korpus[7], jonka perusteella Pauli Saukkonen (1977) laati muun muassa suomen kielen sanojen yleisyystilaston. Turun yliopistossa sähköisiä teksti- ja ääniaineistoja tarjoava Lauseopin arkisto[8] aloitti toimintansa niin ikään 1960-luvulla. Se sisältää sekä puhutun että kirjoitetun kielen aineistoja, kuten kaikki Suomen murrealueet kattavan murrekorpuksen, arkikeskusteluja sisältävän Arkisyn-korpuksen, Mikael Agricolan

teoksia, vieraskielisten kirjoittamaa suomea käsittävän Edistyneiden suomenoppijoiden korpuksen ja akateemiseen suomeen keskittyvän Akateeminen suomi -korpuksen. Monet näistä on morfosyntaktisesti koodattu niin, että korpuksesta tehtävät haut voidaan kohdistaa esimerkiksi kieliopillisena subjektina toimiviin sanoihin. Helsingin yliopistossa oivallettiin kehittyvän tietotekniikan mahdollisuudet kielentutkimuksessa 1980-luvulla, jolloin koottiin noin 1,5 miljoonaa sanaa historiallista englantia sisältävä The Helsinki Corpus of English Texts,[9] joka Brownin ja LOBin tavoin sisältää näytteitä eri tekstilajeista ja joka ulottuu muinaisenglannin kaudelta varhaisuusenglantiin eli noin vuodesta 730 vuoteen 1710. Helsinki Corpus oli ensimmäinen historiallisen englannin kattava korpus, ja alan tutkijat käyttävät sitä edelleen.

Alkuaikojen korpustutkimus liittyi erityisesti variationistiseen lähestymistapaan, jossa pyrittiin selvittämään, kuinka kielen vaihtelua ja muutosta voidaan ymmärtää systemaattisesti ja kuinka näennäisesti sattumanvaraisia kielellisiä valintoja voidaankin selittää erilaisten kielenulkoisten ja -sisäisten muuttujien avulla (ns. *orderly heterogeineity*; Weinreich, Labov & Herzog 1969). Tällaiseen tutkimukseen korpus on välttämätön työkalu, koska se tarjoaa riittävän määrän systemaattisesti järjesteltyä aineistoa, josta voidaan helposti hakea ja laskea kielen piirteiden frekvenssejä. Korpustutkimus yhdistyy edelleenkin vahvasti määrällisiin menetelmiin ja pyrkimykseen selvittää kielen systematiikkaa, mutta nykyisin myös monet tavallisesti laadullisia menetelmiä soveltavat kielentutkimuksen suuntaukset, kuten diskurssintutkimus ja pragmatiikka, hyödyntävät korpuksia.

Viime vuosikymmenen trendejä korpusrintamalla onkin ollut juuri korpusten käytön leviäminen erilaisten tutkimusperinteiden osaksi (ks. Käännöstiede ja sen menetelmät tk.). Lisäksi erilaiset erikoiskorpukset, puhutun kielen korpukset ja aina vain isommat korpukset ovat lisääntyneet, ja myös korpustyökalut, kuten klusteri- ja avainsana-analyysi sekä muut määrälliset menetelmät, ovat kehittyneet tietojenkäsittelytieteen ja korpuslingvistiikan tutkijoiden yhteistyön myötä (ks. Määrällinen korpuslingvistiikka; Historiallinen korpuslingvistiikka tk.).

Suomessa korpusten käyttöä on tukenut ja edistänyt FIN-CLARINin perustaminen 2000-luvulla. FIN-CLARIN on suomalaisten yliopistojen,

CSC:n ja Kotimaisten kielten keskuksen eli Kotuksen muodostama konsortio, jonka tehtävänä on auttaa esimerkiksi kielentutkijoita käyttämään, säilyttämään ja jakamaan aineistoja. FIN-CLARIN ylläpitää Kielipankkia,[10] jonka kautta tutkijat voivat käyttää laajaa valikoimaa valmiita aineistoja ja myös esittää omia aineistojaan liitettäviksi kokoelmiin muiden käytettäviksi.

Tulevaisuuden korpuksissa voisi ajatella eri modaliteettien, kuten tekstin, äänen ja kuvan, sekä metadatan olevan tutkijan helposti saatavilla siten, että niistä voidaan tehdä ristikkäisiä hakuja ja erilaisia visualisointeja esimerkiksi kaavioiden, karttojen tai animaatioiden muodossa. Näin tulevaisuuden korpustyökalu voisi esimerkiksi havainnollistaa automaattisesti vaikkapa tietyn ajankohtaisen sanan, kuten *ihqu* tai *maahanmuutto,* yleistymistä ajan myötä tai sijoittaa kartalle aineistossa havaittujen murresanojen käyttöä.

3. Erilaisia korpuksia

1960-luvun alun jälkeen korpusten koostamisesta on tullut yhä helpompaa ja yleisempää, ja nykyään erilaisia aineistoja on valtavasti saatavilla. Näitä aineistoja voidaan ryhmitellä eri tavoin niiden kokoamisperiaatteiden, käyttötarkoituksen ja niiden sisältämien metatietojen mukaan.

Metatiedot ovat korpukseen konelukuisessa muodossa lisättyjä merkintöjä esimerkiksi korpuksen tekstien kirjoittajista, kirjoitusajankohdasta tai -paikasta tai vaikka puheen prosodiasta. Metatietojen lisäämisestä puhutaan vaihtelevasti joko koodauksena tai annotointina. Tavallinen esimerkki kieliaineistossa on morfologia- tai syntaksiannotoinnit tai -koodaukset, mutta periaatteessa tekstiin voidaan annotoida mitä tahansa tietoja, jotka katsotaan hyödyllisiksi (ks. Määrällinen korpuslingvistiikka; Laadullinen aineistopohjainen kielentutkimus tk.). Oleellista on, että annotoinnit ovat koneellisesti luettavassa muodossa. Näin niitä voidaan käyttää hyväksi koneellisessa analyysissa sekä korpushakuja tehtäessä.

Metatietoja voidaan lisätä joko käsin tai koneellisesti jonkin ohjelman avulla riippuen siitä, mitä metatietoja korpukseen halutaan ja kuinka luotettavia niiden pitää olla. Esimerkiksi korpuksen sanaluokat voidaan monissa tapauksissa tunnistaa automaattisesti riittävällä tarkkuudella. Erityisesti yleiskielelle sanaluokkien tunnistus toimii varsin hyvin, vaikka puheenomaista kieltä tai murteita sisältävien tekstien sanaluokkien tunnistaminen vaatii vielä kehittämistä. Aineistot vaihtelevat sen mukaan, kuinka paljon metatietoa ne sisältävät. Toiset aineistot voivat sisältää hyvin yksityiskohtaiset merkinnät, kun taas toisissa ei ole juurikaan lisättyä tietoa.

Esimerkkinä metatietojen koodaamisesta niin sisällölliseltä kuin tekniseltäkin kannalta voi tutustua avoimesti verkossa saatavilla olevaan Anni Sairion kokoamaan 1700-luvun yksityiskirjeenvaihtoa sisältävään Bluestocking Corpukseen[11], jonka metatiedot löytyvät sekä erillisestä excel-tiedostosta että xml-koodattuina tekstitiedostoista.

Ryhmittelemme alla aineistoja erilaisiin korpustyyppeihin, jotta lukija voi ryhtyä pohtimaan, miten olemassa olevat aineistot voisivat palvella hänen tutkimuksensa tarpeita.

3.1. Yleiskorpukset

Näiden korpusten tavoitteena on kuvata kieltä yleisellä tasolla, ja tavallisesti ne pyritäänkin koostamaan eri lähteistä niin, että ne olisivat mahdollisimman kattavia. Monet näistä korpuksista ovat **kansalliskorpuksia** (engl. *national corpora*). Tämä tarkoittaa, että ne kuvaavat jonkin maan kieltä, kuten vaikka British National Corpus[12] englantia tai Venäjän kansalliskorpus[13] venäjää.

3.2. Synkroniset aineistot

Synkronisia aineistoja käytetään tavallisesti kuvaamaan kielessä ilmenevää vaihtelua jonakin tiettynä ajankohtana. Näiden avulla voidaan vertailla, miten kieltä käytetään esimerkiksi eri alueilla. Tyypillinen esimerkki on The International Corpus of English (ICE)[14], joka on suunniteltu

eri englannin varieteettien tutkimiseen. Se sisältää 20 miljoonan sanan alakorpusta, joista jokainen edustaa jotain englannin maantieteellistä varieteettia, kuten Australia, Hong Kong tai USA. Vertailun helpottamiseksi jokainen alakorpus on koostettu samoilla kriteereillä, ja niissä on samankaltaiset kielioppimerkinnät.

3.3. Diakroniset aineistot

Diakronisia aineistoja käytetään kielen vaihtelun ajalliseen tarkasteluun, eli niistä tutkitaan, miten kieli on muuttunut jonkin ajan kuluessa. Esimerkki diakronisesta korpuksesta on edellä mainittu The Helsinki Corpus of English Texts, joka kattaa noin tuhat vuotta englantia. Diakronisista aineistoista kerrotaan enemmän tämän teoksen luvussa Historiallinen korpuslingvistiikka.

3.4. Puhutun kielen aineistot

Sen lisäksi, että yleiskorpuksissa on omat osionsa puhutulle kielelle, on myös aineistoja, jotka sisältävät ainoastaan puhuttua kieltä. The London-Lund Corpus (LLC)[15] on ensimmäinen elektroninen spontaanin puheen korpus. Se sisältää 500 000 sanaa puhuttua brittienglantia. Aineisto on kirjoitettu puhtaaksi eli litteroitu, ja siinä on mukana varsin yksityiskohtaiset merkinnät tekstin prosodiasta.

Myös Suomessa on kerätty monia puhutun kielen aineistoja. Esimerkiksi yllä jo mainittiin arkikeskusteluja sisältävä Arkisyn ja Lauseopin arkiston murrekorpus, jotka molemmat on kieliopillisesti koodattu eli niihin on merkitty sanojen ja virkkeiden morfologisia ja syntaktisia ominaisuuksia. Lisäksi Lauseopin arkistolla ja Kotimaisten kielten keskuksella on koottuna äänitteitä, joita ei ole kaikkia litteroitu: Kotuksen Suomen kielen nauhoitearkisto sisältää noin 24 000 tuntia äänitteitä suomen murteista, ja Lauseopin arkiston äänitearkiston aineistoissa on muun muassa murrehaastatteluja ja luentojen, kokousten yms. tilaisuuksien nauhoitteita.

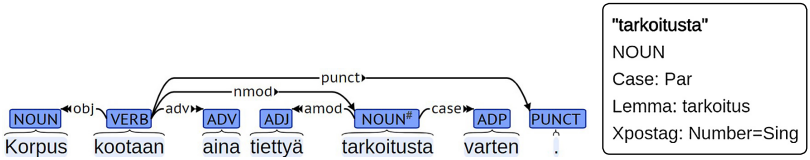
3.5. Puupankit eli syntaksiannotoidut korpukset

Kieliaineistoja, jotka sisältävät yleensä käsin tarkistetut syntaksimerkinnät, sanotaan puupankeiksi. Puupankkeja voidaan käyttää aineistoina kielentutkimuksessa. Erityisesti ne soveltuvat tutkimuksiin, joissa tarkastelun kohteena on jokin syntaktinen ilmiö. Syntaksimerkintöjen ansiosta puupankeista voi helposti hakea näitä ilmiöitä. Lisäksi puupankit ovat tarpeellisia kieliteknologiassa. Koneoppimisen avulla niiden pohjalta voidaan kehittää syntaksijäsentimiä eli tietokoneohjelmia, jotka analysoivat kielen morfologiaa ja syntaksia automaattisesti. Näitä voi hyödyntää korpuksen koostamisessa metatietojen lisäämiseen.

Ensimmäinen korpus tässä kategoriassa oli amerikkalainen, vuonna 1992 julkaistu Penn Treebank. Ensimmäinen suomenkielinen vapaasti saatavilla oleva puupankki on Turku Dependency Treebank (TDT), jonka ensimmäinen versio julkaistiin 2009. Nykyään TDT on liitetty kansainväliseen Universal Dependencies -puupankkikokoelmaan, joka sisältää suuren määrän puupankkeja eri kielille.[16]

Varhaisemmissa puupankeissa virkerakenne on usein merkitty lausekerakennekieliopin keinoin (ks. Tieteen termipankki s.v. lausekerakennekielioppi). Nykyään syntaksi kuvataan yleensä dependenssikieliopin avulla (ks. Tieteen termipankki s.v. dependenssikielioppi). Tässä virkkeen rakenne ilmaistaan sanojen välisten riippuvuuksien keinoin. Virkkeen juuri on tavallisesti finiittiverbi, josta lähtee riippuvuudet muihin sanoihin.

Kuviossa 1 finiittiverbi *kootaan* on virkkeen juuri. Juuresta lähtevät objekti (obj), adverbiaalit (nmod ja adv) ja välimerkkiä merkitsevä punct. Sanasta *tarkoitusta* lähtee vielä adjektiivimääriteriippuvuus (amod) tätä määrittävään adjektiiviin *tiettyä* ja case-riippuvuus postpositioon *varten*. Kuviossa sanojen sanaluokka- ja morfologiatiedot ovat näkyvissä vain sanalle *tarkoitusta*, ja niistä käy ilmi, että sanan sanaluokka on NOUN eli substantiivi, sijamuoto partitiivi, lemma eli perusmuoto *tarkoitus* ja luku yksikkö. Todellisuudessa syntaksiannotointien taustalla puupankeissa on vastaavat tiedot kaikille sanoille.



Kuvio 1. Dependenssisyntaksianalyysi virkkeelle.

3.6. Internet korpuksena ja koneellisesti kootut aineistot

Internetin kehityksen myötä digitaalisessa muodossa on helposti ja nopeasti saatavilla valtavat määrät tekstiä, josta voi koota kieliaineiston. Ensimmäisiä tällaisia koneellisesti internetistä koottuja aineistoja olivat WaCky Corpora -aineistot[17], jotka sisältävät miljardeja sanoja usealla eri kielellä. Näissä aineistot on koottu rajoittamalla tekstihaku maakohtaisiin päätteisiin, kuten .uk tai .fr. WaCky-aineistojen jälkeen vastaavia aineistoja on koottu yhä laajemmalle kielivalikoimalle. Myös suomelle on Turun yliopistossa koottu, miljardeja sanoja sisältävä Finnish Internet Parsebank, joka on käytettävissä samasta internetkäyttöliittymästä kuin yllä mainittu Universal Dependencies -puupankkikokoelma.[18]

Koneellisesti kootuilla aineistoilla on monia etuja: ne ovat erittäin laajoja, ja ne sisältävät valtavasti kielellistä variaatiota myös tekstilajeista, joita käsin koottuihin aineistoihin harvoin päätyy. Toisaalta niiden käyttöön kieliaineistona liittyy myös haasteita: siinä missä käsin koottujen aineistojen sisältö perustuu huolella pohdittuihin kriteereihin, koneellisesti kootuissa aineistoissa otetaan tavallisesti mukaan kaikki tekstit, jotka sisältävät tietyn vähimmäismäärän haluttua kieltä. Tämän takia koneellisesti kootuista aineistoista on hankala sanoa, millaista kieltä ne sisältävät tai mitä ne oikeastaan edustavat. Tämä pätee myös esimerkiksi valtavaan Google Books -aineistoon, jota voi hakea Google Ngram Viewer -ohjelman avulla.[19]

4. Korpuksen kokoaminen

Jos valmiit aineistot eivät sovellu oman tutkimuksen tarpeisiin, tutkija voi päättää koota oman korpuksensa. Korpusta koottaessa on kuitenkin tärkeää muistaa, että korpus ei ole mikä tahansa kokoelma kieltä. Jotta korpus voisi toimia tieteellisen tutkimuksen aineistona, sen on tarkoitus edustaa jotain kieltä tai sen osaa. Tämän takia korpus kootaan aina johonkin tiettyyn tarkoitukseen, ja sen kokoonpanoon vaikuttavat monenlaiset tutkimukselliset ja käytännölliset seikat.

Korpuksen sisältö vaikuttaa tutkimuskysymyksiin, joihin sen avulla voidaan vastata. Esimerkiksi pelkästään radio-ohjelmia sisältävän korpuksen avulla ei voi tarkastella kielen rakenteita puhutussa kielessä ylipäätään, tai pelkästään uutisia käsittävällä aineistolla ei voi vastata kysymyksiin kirjoitetun kielen kehityksestä. Usein puhutaan kriteereistä, joita asiallisesti kerätyn korpuksen tulee täyttää. Esittelemme näitä seuraavissa luvuissa.

Lisäksi korpuksen kokoajan on syytä pohtia ja selvittää jo valmisteluvaiheessa aineiston muotoon, säilytykseen ja jakeluun liittyviä kysymyksiä, sillä nämä eivät ole vain tutkijan harkinnan varaisia seikkoja. Esimerkiksi aineiston säilytyksestä ja jakelusta vastaava infrastruktuuripalvelu voi asettaa aineistolle omia vaatimuksiaan. Aineistoon, sen säilytykseen ja käyttöön voi liittyä myös erilaisia juridisia seikkoja, kuten tekijänoikeus- ja tietosuojakysymyksiä (ks. Johdannon luku 1 tk.). Palaamme näihin lyhyesti tämän luvun lopussa.

4.1. Edustavuus

Nykykieltä tutkittaessa on harvinaista, että tutkimuksen kohteena olevaa kieltä tai sen osaa voitaisiin tutkia kokonaisuudessaan. Esimerkiksi, jos tutkimuksen kohteena on jokin tietty rakenne kirjoitetussa nykysuomessa, korpuksena on mahdotonta käyttää kaikkia tekstejä, joita suomeksi on kirjoitettu. Tämän takia tutkimuksessa käytetty korpus on käytännössä aina otos tutkimuksen kohteena olevasta kielestä tai sen variantista.

Siksi kerätyn aineiston pitää edustaa sitä mahdollisimman hyvin. Ainoastaan tällä tavalla tutkimustulokset voidaan yleistää, eli niiden voidaan ajatella koskevan paitsi tutkimusaineistoa myös sen edustamaa kielivarianttia yleensä.

Edustavan otoksen laatiminen on varsin haastavaa, eikä siihen ole yhtä oikeaa vastausta. Tärkeää on sisällyttää korpukseen mahdollisimman laaja valikoima erilaisia kielenkäyttötilanteita siitä kielestä tai sen osasta, jota korpuksen halutaan edustavan. Tavallisesti edustavuuden asettamat haasteet ratkaistaan koostamalla aineisto alakorpuksista, jotka edustavat jotain tiettyä tekstilajia tai kielenkäyttötilannetta. Esimerkiksi yleiskorpus British National Corpus on hierarkkisesti koostettu niin, että aineisto on ensinnäkin jaoteltu puhuttua ja kirjoitettua kieltä käsittäviin osioihin, jotka taas koostuvat erillisistä alaosioista, kuten faktapohjaiset tekstit aiheesta vapaa-aika (engl. *informative: leisure*) ja faktapohjaiset tekstit aiheesta taide (engl. *informative: arts*). Samankaltaisia jaotteluita käytetään myös niin sanotuissa erityisalojen korpuksissa, kuten vaikka oppijoiden kieltä sisältävissä aineistoissa. Esimerkiksi LAS2 – Edistyneiden suomenoppijoiden korpus sisältää tenttivastauksia, tutkielmatekstejä ja esseitä.

4.2. Koko

Yleinen pohdinnan aihe korpusta koostettaessa on sen koko. Onko suurempi aina parempi? Mikä on riittävää? Onko ylärajaa? Jos aineisto on liian pieni, siitä saatavat tulokset voivat olla vääristyneitä, mutta isokin aineisto voi antaa kummallisen kapean kuvan kielestä. Iso korpus voi olla sillä tavalla vääristynyt, että se on sisällöltään kovin homogeeninen. Esimerkiksi lähinnä tiettyä genreä, kuten vain kaunokirjallisia tekstejä, sisältävä korpus tuskin kuvaa kovin hyvin kielen kirjoa. Toisaalta valtava aineisto voi olla teknisesti hankalakäyttöisempi kuin pieni ja näppärä korpus, jonka voi tallentaa omalle koneelle ja jota voi selata valmiilla korpusohjelmilla.

Aineiston riittävä koko riippuu myös sen käyttötarkoituksesta. Suuresta koosta on hyötyä esimerkiksi sanastontutkimuksessa, sillä monet

sanat ja kollokaatit eli sanojen yhteisesiintymät voivat olla hyvin harvinaisia. Suuri koko on etu myös yleiskorpuksissa, kuten yllä mainitussa BNC:ssä, joiden tavoitteena on kuvata kieltä yleisesti. Jos yleiskorpus on liian pieni, se ei sisällä riittävästi kielen variaatiota ja kuvaa ainoastaan joitain sen yksittäisiä alaluokkia. Toisaalta tarkempiin tutkimuskysymyksiin tai aihepiirin tai tiheästi esiintyvän kielellisen ilmiön tutkimiseen pienempikin aineisto voi olla riittävä.

Kokoa ei myöskään mitata pelkästään aineiston sanamäärällä. Pitää myös pohtia, otetaanko tekstit aineistoon kokonaisuudessaan vai ainoastaan osina niin, että jokaisessa luokassa on tietty sanamäärä. Lisäksi täytyy ottaa huomioon erilaisten tekstien määrä alaluokissa. Aineiston alaluokkien pitää olla tasapainossa niin, että yksikään luokka ei ole liian vallitseva muihin nähden. Esimerkiksi BNC:ssä kirjoitettuja tekstejä on kokonaisuudesta 90 % ja puhuttuja 10 %. Synkroninen eli kieltä tiettynä ajankohtana kuvaava The Longman/Lancaster Corpus[20] taas sisältää 50 % brittienglantia, 40 % amerikanenglantia ja 10 % muita variantteja, kuten australian-, afrikan- ja irlanninenglantia. Tasapaino täytyy pitää mielessä aina yksilötasolle asti, sillä suhteettoman suuri määrä tekstiä yhdeltä kirjoittajalta tai puhujalta voi tuottaa tuloksia, jotka heijastelevat lähinnä tämän kielenkäyttäjän idiosynkraattisia piirteitä.

4.3. Käytännön rajoitteet

Korpuksen koostamiseen liittyy myös monia käytännön rajoitteita, jotka määrittävät lopputulosta eli käytettävää korpusta. Jotkin rajoitteet voivat olla luonnollisia; esimerkiksi historiallisilta aikakausilta ei välttämättä ole saatavilla tiettyjä aineistoja. Toisissa aineistoissa, kuten arkaluonteisia tietoja sisältävissä potilasasiakirja-aineistoissa, saatavuus voi muuten olla rajoitettua. Puhuttuja aineistoja taas voi olla työläs tuottaa.

Ainakin historiallisten kielimuotojen tutkijat ovat kirjoittaneet paljon ”huonon datan” ongelmasta, kun ennen 1900-lukua käytettyjen kielimuotojen tutkimus on perustunut väistämättä ainoastaan kirjallisiin lähteisiin. Koska kielellisten innovaatioiden katsotaan syntyvän epämuodollisessa puheessa, on tätä historiallisessa kielentutkimuksessa

approksimoitu esimerkiksi yksityiskirjeiden ja kirjoitetun puheen, kuten näytelmätekstien, avulla. Lienee kuitenkin parempi käyttää olemassa ja saatavilla olevaa dataa kuin jättää aineistoja kokonaan käyttämättä, vaikka niiden avulla ei voidakaan vastata kaikkiin mahdollisiin tutkimuskysymyksiin.

4.4. Avoin saatavuus

Korpuksen koostaminen on työläs prosessi. Lopputulos kuitenkin palkitsee: valmis, huolella koostettu aineisto tarjoaa arvokasta tietoa kielestä ja sen käytöstä. Jotta mahdollisimman monet voisivat hyötyä tästä jo tehdystä työstä, kannattaa varmistaa, että valmis korpus on tallennettu varmaan paikkaan ja että se on mahdollisimman vapaasti saatavilla. Erityisesti 2010-luvulta lähtien aineistojen vapaata saatavuutta onkin korostettu tutkimuksessa. Esimerkiksi Kielipankin kokoelmiin voi luovuttaa valmiin aineiston muiden käytettäväksi (ks. alla). Korpuksen kokoajan on kuitenkin varmistettava, että hänellä on lupa julkaista aineisto avoimesti kenenkään tekijänoikeuksia tai intimiteettiä loukkaamatta. Lupa-asiat ja mahdollinen aineiston metatietojen keruu kannattaa selvittää jo ennen koostamisen aloittamista. Näistä saa lisätietoa esimerkiksi FIN-CLARINin internetsivuilta.[21]

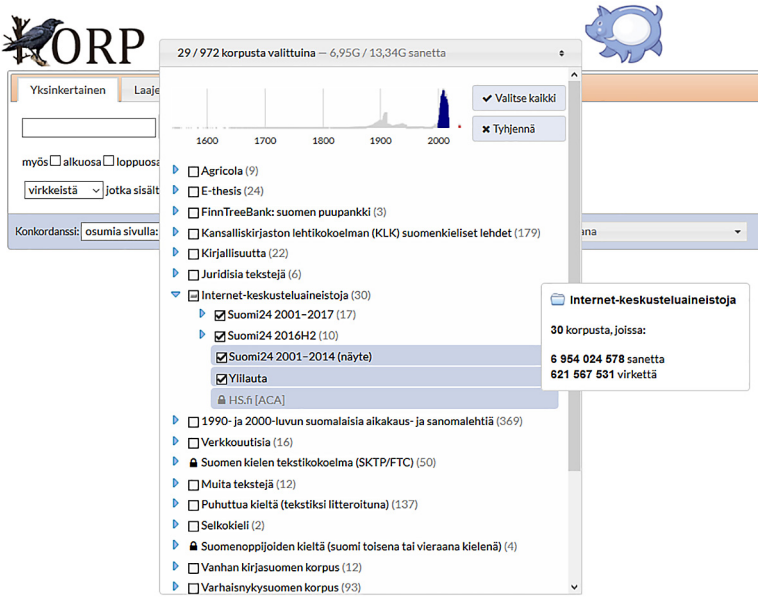
5. Korpustyökaluja

Tutkijan johtotähtenä ovat aina tutkimuskysymykset, olipa aineisto mitä tahansa. Annamme tässä joitakin tyypillisiä esimerkkejä siitä, miten korpuksia voi käyttää, mutta teknistä toteutusta tärkeämpää on miettiä, mitä tietoa kyseiset analyysit tuottavat.

Valmis korpus voi olla käytettävissä erilaisissa muodoissa, kuten tekstitiedostoina tai xml-koodattuina tiedostoina, jolloin tarvitaan vielä erillinen ohjelma tai skriptinkirjoitustaitoja korpushakuja varten.

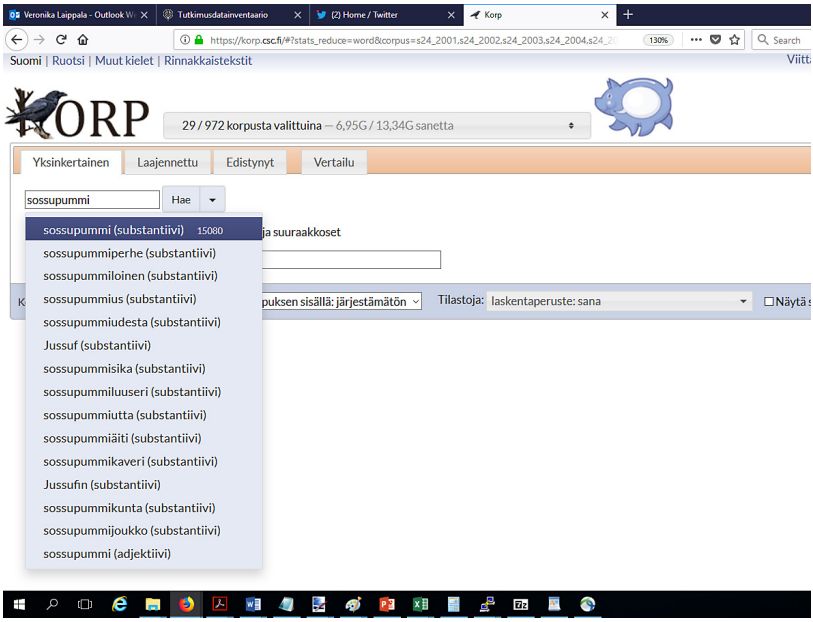
Helpoin tapa useimmille tutkijoille on käyttää korpusta valmiin käyttöliittymän kautta. Esimerkiksi yllä esitellyn Kielipankin aineistoja voi käyttää KORP-hakuliittymän[22] avulla (Borin, Forsberg & Roxendal 2012).

KORP-käyttöliittymä on erittäin monipuolinen ja helppokäyttöinen. Ensimmäiseksi hakua tehdessään käyttäjän tulee valita halutut korpukset monipuolisesta kokoelmasta. Tässä käytämme esimerkkihaussa internetkeskusteluaineistoja, jotka sisältävät yhteensä lähes kolme miljardia sanetta eli sananmuodon esiintymää.



Kuvio 2. Aineiston valitseminen KORPissa.

Aineiston valinnan jälkeen siitä voi tehdä hakuja. Jos aineisto on kieliopillisesti koodattu, haun voi myös kohdistaa tiettyyn sanaluokkaan. Esimerkkihaussa valituissa internetkeskusteluissa näin on, joten ohjelma ehdottaa pudotusvalikossa tarkennuksia hakuun. Haettaessa sanaa *sossupummi* tarkennuksena voi valita esimerkiksi sanaluokan tai jonkin sanan, jonka osana tämä on.



Kuvio 3. Haku KORPissa.

Haun tulokset KORP palauttaa ns. konkordanssinäkymänä. Siinä jokainen hakusana esitetään kontekstissaan omalla rivillään. Hakusanat ovat sijoitettuina allekkain vertaamisen helpottamiseksi.

Kuvio 4. KORPin konkordanssinäkymä.

Kuvio 5. Tekstin ja haetun sanan kuvailutiedot korpuksessa.

Konkordanssinäkymässä jokaisesta rivistä on myös mahdollista nähdä hakusanan alkuperä. Kuvion 4 esimerkin ensimmäisen rivin tiedot kuvataan kuviossa 5. Siitä käy ilmi, että teksti on peräisin Suomi24-aineistosta (Aller Media Oy 2014). Lisäksi kuvataan tarkemmin tekstin piirteitä ja sanan piirteitä. Näiden avulla hakusanasta saa enemmän ja tarkempaa tietoa.

6. Käytä valmiita korpuksia!

Moniin tarkoituksiin sopivia valmiita korpuksia on avoimesti saatavilla, ja erityisesti gradutöitään miettivien opiskelijoiden kannattaa tutustua niihin. Valmis aineisto säästää paljon tutkijan aikaa. Korpuksen metatiedoista voi lukea, millä periaatteilla korpus on koostettu. Aineistona sen pitäisi olla luotettava ja edustava. Sanalla sanoen sen kokoaja on jo ratkaissut monta ongelmaa tutkijan puolesta, ja tutkija voi keskittyä ratkomaan valitsemaansa tutkimuskysymystä. Avoimesti saatavilla oleva aineisto on sikälikin parempi kuin yksittäisen tutkijan tai opiskelijan omaan käyttöön kokoama (pieni) aineisto, että se on myös muiden tutkijoiden käytettävissä, tutkimus on periaatteessa toistettavissa ja sen aukot voidaan osoittaa ja tulosten pohjalta on helpompi suunnitella jatkotutkimuksia.

Vaikka erityisesti jo valmiiksi sähköisessä muodossa olevia tekstejä saattaa olla kohtuullisen helppo koota opinnäytetyön korpukseksi, on kuitenkin korostettava, että huolellisesti kootut laajat korpukset ovat lähes aina ryhmätyön tulosta. Niiden saattaminen valmiiksi ja muiden tutkijoiden saataville kestää vuosia ja vaatii monenlaisia resursseja ja verkostoja, joita yksittäisellä tutkijalla tai opiskelijalla harvemmin on.

Eri kielten tai genrejen osalta valmiiden korpusten saatavuudessa on toki eroja. Englanti on luultavasti maailman tutkituin kieli, joten englannin korpuksia on jo ehditty kokoamaan moneen lähtöön, mikä näkyy tämänkin luvun korpusesimerkeissä, ja lisää esimerkkejä voi hakea Corpus Resource Database (CoRD) -tietokannasta[23]. Kielipankin tarjonta laajentuu jatkuvasti tutkijoiden tallentaessa sinne aineistojaan ja kattaa tällä hetkellä monien suomen kielen korpusten lisäksi useita muita kieliä ja tekstilajeja, kuten Pohjoissaamen korpus, Lettere amorose ja Lyydin korpus.[24] Joistain kielistä valmiita korpuksia ei juurikaan ole, ja olisikin syytä miettiä, voitaisiinko tällaisten kielten tutkimusta edistää kokoamalla avoimesti saatavilla olevia korpuksia.

Valmistakaan aineistoa ei voi käyttää sokkona, vaan sen kokoamisperiaatteisiin on tutustuttava huolella. Ensin kannattaa kaivaa esiin valitun korpuksen kokoamisperiaatteita ja sisältöä selostava käsikirja

eli korpusmanuaali, jos korpuksen kokoajat ovat sellaisen laatineet. Korpusmanuaali on tutkijan keskeinen tietopaketti aineistosta, ja myös oman korpuksen kokoajan on syytä dokumentoida aineistonsa luonne paitsi omaan käyttöön myös mahdollisia muita käyttäjiä varten. ICAME Corpus Manuals -sivustolla[25] voi tutustua erilaisiin korpusmanuaaleihin ja niiden sisältämään tietoon. Lisäksi kannattaa tutustua aiempaan kyseisestä korpuksesta tehtyyn tutkimukseen, jota voi aluksi hakea vaikkapa Google Scholarin avulla, ellei sitä ole koottu yhteen esimerkiksi korpuksen verkkosivulle.

Kävimme edellä läpi joitakin korpuksen kokoamiseen liittyviä yleisiä periaatteita, jotka on hyvä huomioida paitsi oman korpuksen kokoamisessa myös valmiin korpuksen käytössä. Hyvä korpusmanuaali ottaa kantaa juuri edellä kuvattuihin seikkoihin ja esittää kokoamisperiaatteet läpinäkyvästi.

7. Korpustutkimuksen rajoitteet ja ongelmat

Ehkä aivan ensimmäinen harha, jonka valtaan korpustutkija saattaa huomaamattaankin joutua, on kuvitella, että korpuksessa on kielen koko kirjo ja ”lopullinen totuus” kielestä. Suomalaisen korpustutkimuksen uranuurtajiin lukeutuva Matti Rissanen (1989) viittaa tähän tutkijan harhaan käsitteellä *God's Truth Fallacy*. Korpus ei ole koko totuus kielestä, vaan laajoihinkin korpuksiin voi liittyä monenlaisia edustavuuden rajoitteita. Korpukset ovat esimerkiksi paljolti olleet kirjoitetun kielen aineistoja, koska puhetta on paljon vaikeampi kerätä ja muuntaa tekstimuotoon, ja puhekin on saattanut olla tietyntyyppistä suunniteltua tai harkittua puhetta pikemminkin kuin arkipuhetta. Aineisto on usein sitä, mitä saa kohtuullisella vaivalla. On siis täysin mahdollista, että kielellinen ilmiö on olemassa, vaikka sen esiintymiä ei korpuksesta löytyisikään.

Korpuksista ei voi myöskään hakea kaikkea mahdollista automaattisesti, ja haettavuus vaikuttaa siihen, mitä ja miten voidaan tutkia. Sanoja

ja sanan osia voi hakea, ja jos kyseessä on syntaksiannotoitu korpus, kenties sanaluokat tai lauseenjäsenet ovat helposti tutkijan ulottuvilla. Myös semanttisia luokitteluja saattaa sisältyä korpukseen, tai tutkijan on mahdollista luoda niitä itse semanttisten taggereiden eli semanttisia rooleja tunnistavien ohjelmien avulla. Vaikka tällaiset merkinnät helpottavat monella tapaa tutkijan työtä, on kuitenkin tiedostettava, että mitä enemmän hyödyntää valmiita luokitteluja, sitä sidotumpi on tiettyyn kielioppimalliin ja sen rajoitteisiin (ks. Kieliopin tutkimus ja kielioppiteoriat tk.).

Jotkin kielen ilmiöt ovat kerta kaikkiaan vaikeammin haettavia kuin toiset. Niin kauan kuin ilmiön voi sitoa tiettyihin sanoihin tai fraaseihin, hakeminen on helppoa. Pragmaattisia ja vuorovaikutteisia ilmiöitä voi kuitenkin olla vaikea tutkia korpusmenetelmin, jollei niitä pysty operationaalistamaan haettavalla tavalla eli yhdistämällä niitä sanoihin tai fraaseihin.

Korpustutkijat kuvaavat mielellään omaa tutkimustaan sloganilla ”tutkimus alkaa siitä, mihin laskeminen loppuu” (engl. *research begins where counting ends*). Tämä lentävä lause kuvaa hyvin sitä, että laskeminen ei ole korpustutkimuksen itsetarkoitus vaan laskemalla pystytään vastaamaan tutkimuskysymyksiin. Korpustutkimuksen laskennallisuuteen saattaa kuitenkin liittyä myös ongelmia, joista on hyvä olla tietoinen. Kun lingvistiseen koulutukseen ei läheskään aina sisälly edes alkeellista tilastotieteellistä elementtiä, korpustutkijan on erikseen perehdyttävä alansa laskennalliseen puoleen ja kenties värvättävä projekteihinsa alan osaajia. Jotkin korpusohjelmat, kuten WordSmith[26] ja AntConc[27], tarjoavat tilastollisiin menetelmiin perustuvia työkaluja, jolloin tutkijan ei välttämättä tarvitse itse laskea, mutta hänen on kuitenkin pääpiirteissään ymmärrettävä, mitä työkalut hänen puolestaan tekevät.

Korpusten käyttö mahdollistaa kielen tarkastelun erillään kielenkäytön tilanteesta, mutta tutkijan on aina tunnettava korpuksen tekstit ja niiden kielellinen ja kulttuurinen konteksti. Esimerkiksi koneellisesti internetistä koostettuihin aineistoihin voi olla päätynyt osia, jotka vääristävät analyysia, jos kontekstiin ei kiinnitä huomiota: esimerkiksi virke *Lue lisää* esiintyy hyvin usein internetsivuilla muttei silti kerro juuri mitään siellä käytettävästä kielestä. Kielenkäytön kontekstiin tutustuminen voi aineistosta ja tutkimuskysymyksistä riippuen tarkoittaa esimerkiksi korpuksen sisältämien genrejen ominaispiirteisiin ja tuotantoprosesseihin

perehtymistä tai tekstien kirjoittajien henkilöhistorioiden ja keskinäisten verkostojen kartoittamista ja sijoittamista tutkittavan ajan yhteiskunnalliseen kontekstiin. Tässä kontekstualisointityössä tarvitaan yleensä apuna muiden (alojen) tutkijoiden tuottamaa tietoa, joka voi olla luonteeltaan esimerkiksi historiallista tai yhteiskuntatieteellistä.

8. Yhteenveto

Tässä artikkelissa olemme käsitelleet korpusten käyttöä ja niiden kokoamista kielentutkimuksessa. Olemme myös esitelleet erilaisia korpuksia ja pohtineet korpustutkimuksen hyötyjä ja mahdollisia sudenkuoppia. Kuten mainitsimme johdannossa, korpusten kokoaminen ja käyttö liittyvät läheisesti yhteen tietotekniikan kehityksen kanssa. Jo nyt kielentutkijoilla on tietokoneiden avulla mahdollisuus päästä käsiksi valtaviin aineistoihin, ja automaattiset ohjelmat helpottavat tutkijan tekemää analyysia esimerkiksi syntaksijäsennyksen avulla. Parhaimmillaan menetelmien kehittyminen synnyttää uudenlaisia tutkimuskysymyksiä ja avaa kieleen sellaisia näkökulmia, joita emme vielä osaa edes ennustaa. Tulevaisuudessa onkin mielenkiintoista päästä seuraamaan, miten ala kehittyy.

Aiheesta lisää:

Baker, Paul, Hardie, Andrew & McEnery, Tony. 2006. *Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Biber, Douglas & Reppen, Randi (toim.). 2015. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press. Saatavissa: https://doi.org/10.1017/CBO9781139764377.

Lüdeling, Anke & Kytö, Merja (toim.). 2008. *Corpus Linguistics. An International Handbook*. Osa 1. Berlin & New York (NY): Mouton de Gruyter. Saatavissa: https://doi.org/10.1515/9783110211429.

Lüdeling, Anke & Kytö, Merja (toim.). 2009. *Corpus Linguistics. An International Handbook*. Osa 2. Berlin & New York (NY): Mouton de Gruyter. Saatavissa: https://doi.org/10.1515/9783110213881.2.

McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press. Saatavissa: https://doi.org/10.1017/CBO9780511981395.

VIITTEET

1 http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/.
2 http://www.helsinki.fi/varieng/CoRD/corpora/LOB/.
3 http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/index.html.
4 http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/index.html.
5 http://www.helsinki.fi/varieng/CoRD/corpora/BLOB-1931/.
6 http://www.helsinki.fi/varieng/CoRD/corpora/B-BROWN/.
7 https://www.kielipankki.fi/aineistot/oulu/.
8 https://www.utu.fi/fi/yliopisto/humanistinen-tiedekunta/suomen-kieli-ja-suomalais-ugrilainen-kielentutkimus/lauseopin-arkisto.
9 http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/.
10 http://kielipankki.fi/.
11 http://bluestocking.ling.helsinki.fi/.
12 http://www.natcorp.ox.ac.uk.
13 https://ruscorpora.ru/.
14 http://www.ucl.ac.uk/english-usage/projects/ice.htm.
15 http://www.helsinki.fi/varieng/CoRD/corpora/LLC/.
16 https://universaldependencies.org/
17 http://wacky.sslmit.unibo.it/doku.php?id=start.
18 https://universaldependencies.org/
19 https://books.google.com/ngrams.
20 http://www.pearsonlongman.com/dictionaries/corpus/lancaster.html.
21 https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/FinClarinEsittely.
22 https://korp.csc.fi/.
23 http://www.helsinki.fi/varieng/CoRD/index.html.
24 https://www.kielipankki.fi/aineistot/.
25 http://clu.uni.no/icame/manuals/.
26 http://www.lexically.net/downloads/version6/HTML/index.html?getting_started.htm.
27 http://www.laurenceanthony.net/software/antconc/.

KIRJALLISUUS

Aller Media Oy. 2014. The Suomi 24 Sentences Corpus (2016H2). [tekstikorpus]. Kielipankki. Saatavissa: http://urn.fi/urn:nbn:fi:lb-2017021505.

Borin, Lars, Forsberg, Markus & Roxendal, Johan. 2012. Korp – the corpus infrastructure of Språkbanken. Saatavissa: https://spraakbanken.gu.se/eng/publikationer/korp-%e2%80%93-corpus-infrastructure-spr%c3%a5kbanken.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13, 16–19.

Saukkonen, Pauli. 1977. *Nykysuomen saneiston yleisyystilastoa saneenloppuisessa aakkosjärjestyksessä*. Oulu: Oulun yliopisto.

Tieteen termipankki. 2019. [verkkoaineisto]. [viitattu 28.2.2019]. Saatavissa: https://tieteentermipankki.fi.

Weinreich, Uriel, Labov, William & Herzog, Martin. 1969. Empirical foundations for a theory of language change. Julkaisussa: Lehmann, W. P. & Malkiel, Yakov (toim.) *Directions for Historical Linguistics*. Austin: University of Texas Press, 97–188.

Korpusaineistoja

Esimerkkejä kokoavista metatietokannoista:

- CLARIN Language Resource Inventory: https://www.clarin.eu/content/language-resource-inventory
- Corpus Resource Database (CoRD): http://www.helsinki.fi/varieng/CoRD/
- ICAME Corpus Manuals: http://clu.uni.no/icame/manuals/
- Linguistic Data Consortium: https://www.ldc.upenn.edu
- META-SHARE: http://www.meta-share.org

Keskeisiä monia kieliä sisältäviä korpuskokoelmia:

- CQPweb, Lancasterin yliopiston korpuskokoelma: https://cqpweb.lancs.ac.uk/
- IntraText: http://www.intratext.com/
- Kielipankki: https://www.kielipankki.fi
- Sähköisten kirjojen projekti Gutenberg, ”Project Gutenberg”: http://www.gutenberg.org/
- WaCky-aineistot: http://wacky.sslmit.unibo.it/doku.php?id=start

Keskeisiä suomenkielisiä korpuksia:

- Finnish Internet Parsebank: https://turkunlp.org/finnish_nlp.html#parsebank
- Kansalliskirjaston lehtikokoelma: https://digi.kansalliskirjasto.fi
- Kotimaisten kielten keskuksen sähköiset aineistot: https://www.kotus.fi/aineistot/tietoa_aineistoista/sahkoiset_aineistot_kootusti
- Lauseopin arkisto: https://www.utu.fi/fi/yliopisto/humanistinen-tiedekunta/suomen-kieli-ja-suomalais-ugrilainen-kielentutkimus/lauseopin-arkisto
- Vanhan kirjasuomen korpus: http://kaino.kotus.fi/korpus/vks/meta/vks_coll_rdf.xml

Keskeisiä englanninkielisiä korpuksia:

- British National Corpus: http://www.natcorp.ox.ac.uk/
- Brown Corpus: http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/
  - B-Brown: http://www.helsinki.fi/varieng/CoRD/corpora/B-BROWN/
  - B-LOB-1931: http://www.helsinki.fi/varieng/CoRD/corpora/BLOB-1931/
  - F-LOB : http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/
  - Frown Corpus: http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/
  - Lancaster-Oslo-Bergen (LOB) Corpus: http://www.helsinki.fi/varieng/CoRD/corpora/LOB/
- Corpora of Early English Correspondence (CEEC): http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
- Corpus of Contemporary American English (COCA): https://www.english-corpora.org/coca/
- Corpus of Historical American English (COHA): https://www.english-corpora.org/coha/
- Corpus of Late Modern English Texts v.3.0 (CLMET3.0): https://perswww.kuleuven.be/~u0044428/clmet3_0.htm
- Early English Books Online (EEBO): https://www.english-corpora.org/eebo/

- ECCO-TCP Corpus: http://www.textcreationpartnership.org
- Helsinki Corpus of English Texts (HC): http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/
- International Corpus of English (ICE): http://www.ice-corpora.uzh.ch/en.html
- The London-Lund Corpus (LLC): http://www.helsinki.fi/varieng/CoRD/corpora/LLC/
- Penn Parsed Corpus of Modern British English (PPCMBE2): http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1

Keskeisiä espanjankielisiä korpuksia:

- ADESSE: http://adesse.uvigo.es/
- Corpus Biblia Medieval: http://corpus.bibliamedieval.es/
- Corpus del español actual: http://cea.spanishfn.org/cea/
- Mark Davies - Corpus del español: http://www.corpusdelespanol.org/x.asp (Käyttääkseen tätä korpusta on rekisteröidyttävä, mutta käyttö on periaatteessa ilmaista.)
- Preseea: http://preseea.linguas.net/Corpus.aspx
- Real Academia Española
  - Corpus del español del siglo XXI (CORPES XXI): http://web.frl.es/CORPES/
  - Corpus de referencia del español actual (CREA) : http://corpus.rae.es/creanet.html
  - Corpus diacrónico del español (CORDE): http://corpus.rae.es/cordenet.html
  - Corpus del Nuevo Diccionario Histórico del Español (CDH): http://web.frl.es/CNDHE/view/inicioExterno.view

Keskeisiä italiankielisiä korpuksia:

- Biblioteca della letteratura italiana: http://www.letteraturaitaliana.net/
- Corpus PAISÀ: http://www.corpusitaliano.it/en/
- Corpus di italiano scritto, CORIS/CODIS: http://corpora.dslo.unibo.it/coris_ita.html

- Listaus italiankielisiin korpuksiin: http://www.accademiadellacrusca.it/it/link-utili/banche-dati-dellitaliano-scritto-parlato

Keskeisiä ranskankielisiä korpuksia:
- Corpus Frantext: https://www.frantext.fr
- Corpora of Computer-Mediated Communication CoMeRe: https://repository.ortolang.fr/api/content/comere/v3.3/comere.html
- Korpuskokoelma Ortolang: http://www.cnrtl.fr/

Keskeisiä ruotsinkielisiä korpuksia:
- Oslon yliopiston pohjoismaalainen murreaineisto: http://www.hf.uio.no/iln/tjenester/kunnskap/sprak/korpus/talespraksk orpus/nordisk-dialekt/index.html
- Suomenruotsalainen murrekorpus TALKO: http://www.sls.fi/talko
- Språkbanken: https://spraakbanken.gu.se/

Keskeisiä saksankielisiä korpuksia:
- Clarin-D-keskusten korpusesittely: https://www.clarin-d.net/en/corpora
- Itävallan Akatemian korpus: http://www.aac.ac.at/
- Katsaus saksankielisiin korpuksiin: Teubert, Wolfgang. 2012. Corpora: German language. Julkaisussa: *The Encyclopedia of Applied Linguistics*. Wiley Online Library, 1–10.
- Puhutun saksan arkisto ”Archiv für gesprochenes Deutsch”: http://agd.ids-mannheim.de/index.shtml
- Saksan kielen referenssikorpus ”Deutsches Referenzkorpus” (DeReKo): http://www1.ids-mannheim.de/kl/projekte/korpora.html
- Saksan kielen digitaalisen sanakirjan (DWDS) korpukset: https://www.dwds.de/
- Sveitsinsaksan korpus ”Schweizer Textkorpus”: https://www.chtk.ch/index.php/de/

- Virheannotoitu oppijansaksan korpus – ”Falko, fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache”: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko

Keskeisiä vironkielisiä korpuksia:

- Eesti Keeleressursside Keskus: https://keeleressursid.ee/et/keeleressursid/tekstikorpused.
- Eesti Kirjakeele Korpus 1890–1990: http://www.cl.ut.ee/korpused/baaskorpus
- Eesti keele koondkorpus: https://www.keeletehnoloogia.ee/et/ekktt/ekktt-projektid/eesti-keele-koondkorpus/koondkorpus
- Eesti vahekeele korpus (oppijankielen korpus): http://evkk.tlu.ee
- Tasakaalus korpus: http://www.cl.ut.ee/korpused/grammatikakorpus

Esimerkkejä muunkielisistä korpuksista ja kokoelmista:

- Komi: http://wiki.fu-lab.ru/index.php/%D0%92%D0%B8%D0%BA%D0%B8_FU-Lab
- Portugali: Davies & Ferreira (2006), historiallisen portugalin korpus: https://www.corpusdoportugues.org
- Saame: Tromssan yliopiston saamen kieliteknologia: http://giellatekno.uit.no/index.fin.html
- Tataari: Corpus of Written Tatar: http://www.corpus.tatar/en Tatar National Corpus ”Tugan tel”: http://web-corpora.net/TatarCorpus/search/?interface_language=en
- Turkki: Turkish National Corpus: https://www.tnc.org.tr/
- Udmurtti: http://web-corpora.net/UdmurtCorpus/search/index.php?interface_language=en
- Unkari: Unkarin kielen kansalliskorpus: http://corpus.nytud.hu/mnsz/index_eng.html
- Venäjä: Venäjän kansalliskorpus https://ruscorpora.ru/

Esimerkkejä käännöstieteellisen korpustutkimuksen keskeisistä korpuksista:

- Community Interpreting Database (ComInDat): http://www.yorku.ca/comindat/comindat.htm
- The English Norwegian Parallel Corpus (ENPC): http://www.hf.uio.no/ilos/english/services/omc/enpc/
- The European Parliament Interpreting Corpus (EPIC): http://metashare.elda.org/repository/browse/european-parliament-interpretation-corpus-epic/fde33884de7611e2b1e-400259011f6ea48ac8ce0b41f48e6be224bbee9d59cb9/
- Käännössuomen korpus. Ks. esim. Mauranen, Anna. 2000. Strange strings in translated language. A study on corpora. Julkaisussa: Olohan, Maeve (toim.) *Intercultural faultlines. Research models in translation studies I. Textual and cognitive aspects*. Manchester: St. Jerome, 119–141.
- The Open Parallel Corpus Opus: http://opus.lingfil.uu.se/
- Translational English Corpus: https://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-english-corpus-tec/
- YLE-korpus: https://wiivi.uef.fi/crisyp/disp/_/fi/cr_redir_all/fet/fet/sea?direction=1&id=-29931

Esimerkkejä kielitypologian keskeisistä aineistoista:

- AUTOTYP-tietokanta: https://github.com/autotyp/autotyp-data
- The Ethnologue: https://www.ethnologue.com/
- Glottolog: https://glottolog.org
- Typological Database System: https://languagelink.let.uu.nl/tds/index.html
- The World Atlas of Language Structures Online: https://wals.info/
- Cross-Linguistic Linked Data: https://clld.org/
- Database of Places, Language, Culture and Environment: https://d-place.org/